Untersuchungen zur Mikroevolution von Methicillin-resistenten *Staphylococcus aureus* im klonalen Komplex CC5

Von der Fakultät für Lebenswissenschaften
der Technischen Universität Carolo-Wilhelmina zu Braunschweig
zur Erlangung des Grades einer
Doktorin der Naturwissenschaften (Dr. rer. nat.)
genehmigte
Dissertation

von Janina Dordel
aus Göttingen

1. Referent: Prof. Dr. Dieter Jahn
2. Referent: Privatdozent Dr. Ulrich Nübel
eingereicht am: 05.03.2012
mündliche Prüfung (Disputation) am: 16.05.2012

Druckjahr 2012

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

Nübel U, Dordel J, Strommenger B, Westh H, Shukla SK, Zemlicková H, Leblois R, Wirth T, Jombart T, Balloux F & Witte W (2010). A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant *Staphylococcus aureus*. *PLoS Pathogens* **6**(4): e1000855.

Tagungsbeiträge

Dordel J, Boye K, Bartels MD, Westh H, Witte W & Nübel U (2011). MRSA genome variation among and within individual patients. - Vortrag von U. Nübel bei der 63. Jahrestagung der Deutschen Gesellschaft für Hygiene und Mikrobiologie (DGHM), Essen.

Dordel J, Dabrowski PW, Nitsche A, Witte W & Nübel U (2011). Genetic flux and temporal process of diversification in prophages of methicillin-resistant *Staphylococcus aureus*, CC5. - Poster bei der 5. ProkaGENOMICS (European Conference on Prokaryotic and Fungal Genomics), Göttingen.

Dordel J, Dabrowski PW, Nitsche A, Witte W & Nübel U (2010). Dynamics of prophages mapped onto the genome based phylogeny of MRSA, clonal complex 5. - Vortrag bei der 9. IMMEM (International Meeting on Microbial Epidemiological Markers, Wernigerode.

Dordel J, Dabrowski PW, Nitsche A, Witte W & Nübel U (2010). Footprints of selection and the gain and loss of genetic information in genomes from methicillin-resistant *Staphylococcus aureus* clonal complex 5. - Kurzvortrag bei der 3. Gemeinsamen Jahrestagung der Deutschen Gesellschaft für Hygiene und Mikrobiologie (DGHM) und der Vereinigung für Allgemeine und Angewandte Mikrobiologie (VAAM), Hannover.

Dordel J, Witte W & Nübel U (2009). Microevolution of *Staphylococcus aureus* clonal complex 5 investigated on the basis of whole genome sequences. - Poster bei der 4. ProkaGENOMICS (European Conference on Prokaryotic and Fungal Genomics), Göttingen.

Inhaltsverzeichnis

Zusammenfassung

Staphylococcus aureus ist ein weitverbreitetes pathogenes Bakterium, das eine Vielzahl verschiedener Krankheiten auslösen kann. Eine besondere Bedrohung geht von Methicillin-resistenten *S. aureus* (MRSA) aus, die Resistenzen gegen ein weites Spektrum von Antibiotika entwickelt haben. In der vorliegenden Arbeit wurden neuartige Sequenziertechnologien verwendet, um durch die Sequenzierung ausgewählter MRSA-Genome neue Einblicke in die Evolution und die räumliche Ausbreitung dieses Keims zu gewinnen.

Der Vergleich von 24 MRSA-Genomen des klonalen Komplexes CC5 zeigt wenig Variabilität im Kerngenom, aber eine hohe Diversität in mobilen genetischen Elementen, insbesondere Prophagen. Auf der Basis Genom-weiter SNPs wurde die Phylogenie des klonalen Komplexes CC5 mit der zurzeit größtmöglichen Auflösung rekonstruiert. Durch die nachfolgende Abbildung der Prophagen auf die Phylogenie der Wirts-Staphylokokken konnte gezeigt werden, dass der Austausch von Phagensequenzen eine wichtige Rolle für die kurzzeitige Evolution von MRSA spielt. Durch homologe Rekombination akkumulieren die Prophagen offenbar kontinuierlich Sequenzunterschiede, so dass ihre Diversität mit der phylogenetischen Distanz zwischen den Wirtsbakterien zunimmt.

In dieser Arbeit konnte erstmals gezeigt werden, dass MRSA-Genome genügend Informationen enthalten, um auch für Epidemiestämme mit weiter geographischer Verbreitung mit hoher Prävalenz die Übertragung des Erregers zwischen einzelnen Patienten während eines Krankenhaus-Aufenthalts nachzuweisen. Darüber hinaus konnte durch eine phylogenetische Analyse sogar die Richtung der Transmission von einem Patienten auf einen anderen bestimmt werden. Weiterhin konnte die Evolution von MRSA innerhalb einzelner Patienten beobachtet werden. Die Isolate zeigten zwar eine über die Zeit ansteigende, genetische Distanz zu dem jeweils infizierenden Stamm, jedoch nahm gleichzeitig ihre Diversität nicht zu. Vielmehr war jeder Patient zu jedem Zeitpunkt mit einem einzelnen, genetisch homogenen Klon besiedelt.

Sequenzen ganzer Genome ermöglichen es, die Evolution von Pathogenen mit hoher Auflösung zu untersuchen. In naher Zukunft wird die Sequenzierung ganzer Genome eine wichtige Rolle in der mikrobiellen Diagnostik spielen.

Summary

Staphylococcus aureus is a major pathogen that causes a variety of infections. A particular danger originates from methicillin-resistant *S. aureus* (MRSA) which have become resistant to a wide spectrum of antibiotics. For the present work novel sequencing technologies were applied to sequence selected MRSA genomes to get deeper insights into the evolution and the spatial spread of this particular pathogen.

The comparison of 24 MRSA genomes of the clonal complex CC5 revealed little variability within the core genome itself but a high level of diversity within mobile genetic elements, especially prophages. On the basis of genome wide SNPs the phylogeny of the clonal complex CC5 was reconstructed to the highest possible degree. Through the mapping of the prophages onto the phylogenetic tree of the host Staphylococci it could be shown that the exchange of phage sequences plays an important role in the short-term evolution of MRSA. Prophages accumulate sequence differences continuously through homologous recombination, causing an increase in their diversity with the phylogenetic distance between the host bacteria.

This work was able to show for the first time that genome sequences from highly prevalent, epidemic MRSA contain enough information to verify the transmission of the pathogen between patients during hospitalisation. The use of phylogentic methods was successful in determining the direction of transmission from one patient to another. Furthermore, the evolution of MRSA within unique patients was investigated. The isolates showed an increasing genetic distance to the infecting strain but at the time their diversity did not rise. In fact each patient was colonised with a unique genetically homogeneous clone.

Whole genome sequencing provides a detailed insight into the evolution of a pathogen at a high resolution. In the near future the sequencing of whole genomes will play a major role in microbial diagnostic.

1 Einleitung

1.1 *Staphylococcus aureus*

Der schottische Chirurg und Mikrobiologe Alexander Ogston entdeckte 1881 in einer Eiterprobe zwei Mikrokokken: die bereits bekannten - in Ketten organisierten - Streptokokken sowie ein anderes Kokken-Genus, das in Traubenform vorlag.
Ogston nannte sie nach dem griechischen Wort für Weintraube („*staphylè*“) *Staphylococcus*. 1884 isolierte und kultivierte der Mediziner Friedrich J. Rosenbach ebenfalls aus Eiter Staphylokokken und benannte sie - aufgrund ihrer gelben Pigmentierung - *Staphylococcus aureus*. Mehr als ein Jahrhundert später zählt dieser Organismus nach wie vor zu den erfolgreichsten Humanpathogenen.

Systematik. *Staphylococcus aureus* gehört neben 69 weiteren Spezies und Subspezies (Stand: September 2011, www.dsmz.de) der Gattung *Staphylococcus* an, die zur Familie der Staphylococcaceae gezählt werden. Diese zeichnet sich durch eine fakultativ anaerobe Lebensweise aus. Eine Ausnahme bildet die Subspezies *Staphylococcus aureus* subsp. *anaerobius*, die obligat anaerob wächst und an Schafe adaptiert ist.

Morphologie und Eigenschaften. *S. aureus* ist ein 0,8 bis 1,2 μm kleines, kugelförmiges Bakterium. Es besitzt einen Gram-positiven Zellwandaufbau und ist unbeweglich. Obwohl *S. aureus* keine Sporen bildet, ist es aufgrund von pH-Toleranz und einer Resistenz gegen Austrocknung relativ unempfindlich. Somit kann es fakultativ anaerob unter verschiedensten Umweltbedingungen wachsen, wobei Temperaturen zwischen 30 und 37 °C aber bevorzugt werden.

Verbreitung. *S. aureus* ist weltweit verbreitet und gehört zur Resident- und Transientflora des Menschen und anderer Säugetiere und Vögel. Es werden vor allem die Haut, die oberen Atemwege (z.B. Nase) sowie Schleimhäute und die Leistengegend besiedelt, wobei die Kolonisation in den meisten Fällen symptomlos ist (Hartmann 1978, Valle *et al.* 1991, Kloos & Lambe 1991). In Deutschland ist etwa ein Drittel der Bevölkerung mit *S. aureus* besiedelt (von Eiff *et al.* 2001).

Diagnostik. Die Diagnose von *S. aureus* kann aus unterschiedlichem Probenmaterial (z.B. Abstriche, Blut) erfolgen. Die mikroskopische Betrachtung eines Gram-Präparats gibt einen ersten Hinweis auf Gram-positive Kokken. Da Staphylokokken Katalase-Bildner sind, können sie mit Hilfe eines Katalase-Tests von Katalase-negativen Streptokokken unterschieden werden. Ein anschließender Koagulase-Test differenziert zwischen Koagulase-positiven (*S. aureus*, *S. intermedius*) und Koagulase-negativen (*S. epidermis* u.a.) Staphylokokken.

Krankheitsbilder. Obwohl *S. aureus* ein Kommensal der Normalflora ist, besitzt es eine Vielzahl von Pathogenitäts- und Virulenzfaktoren (siehe Tabelle 1.1) und ist unter bestimmten Bedingungen in der Lage, eine große Anzahl verschiedenster Krankheiten auszulösen. Diese lassen sich in drei Gruppen einteilen:

1. lokale, oberflächliche Infektionen - Haut- bzw. Wundinfektionen, Abzesse, Emphyseme.

2. invasive, systemische Infektionen - Osteomylitis, Pneumonie, Endokarditis, Mastitis, Sepsis.

3. toxinvermittelte Erkrankungen - „*Staphylococcal Scalded Skin Syndrom*“ (SSSS), Toxisches Schock Syndrom (TSS), Impetigo contagiosa, Lebensmittelvergiftung.

Eine Infektion mit Staphylokokken umfasst dabei die Schritte Kolonisation, Lokalinfektion, systemische Streuung und/oder Sepsis, Absiedlung und die Toxinose (Lowy 1998).
Vor allem bereits immunsupprimierte Menschen sind anfällig für Infektionen, weshalb *S. aureus* besonders in Krankenhäusern eine Gefahr darstellt und dort für einen Großteil der nosokomialen Infektionen verantwortlich ist. Neben der vereinfachten Übertragung zwischen Patienten oder Patient und Krankenhauspersonal spielt hier vor allem die Fähigkeit zur Bildung von Biofilmen eine wichtige Rolle, durch die Katheter und Implantate kolonisiert werden können. Ein besonderes Problem stellen gegen Antibiotika resistent gewordene Stämme dar, da sie eine erfolgreiche Therapie erschweren. Auf die Methicillin-resistenten *S. aureus* (MRSA) wird in Kapitel 1.2 noch näher eingegangen.

Pathogenitäts- und Virulenzfaktoren. *S. aureus* besitzt eine große Anzahl an Pathogenitäts- und Virulenzfaktoren, die es dem Bakterium ermöglichen, die Immunabwehr des Wirts auf verschiedene Wege zu umgehen und so eine Vielzahl an pathogenen Prozessen und damit Infektionen auszulösen. Tabelle 1.1 enthält eine Übersicht der verschiedenen Faktoren sowie ihrer Wirkung.

Therapie. Bei einer *S. aureus*-Infektion sollte im Vorfeld ein Antibiogramm erstellt werden, um ein angemessenes Antibiotikum auswählen zu können. Ungefähr 80 % aller *S. aureus* sind durch das *blaZ*-Gen Penicillin-resistent. Zur Behandlung von Infektionen können entweder Penicillase-feste Penicilline („Staphylokokkenpenicilline“), eine Kombination mit β-Lactamase-Hemmern oder Cephalosporine verabreicht werden. Wird eine Resistenz gegen Oxacillin festgestellt, besitzt *S. aureus* das *mecA*-Gen und ist damit gegen alle β-Lactame resistent. Die Therapie erfolgt je nach Antibiogramm; liegt noch keines vor, können Reserveantibiotika gegeben werden (siehe Kapitel 1.2 *Therapie*).

Tabelle 1.1: Pathogenitäts- und Virulenzfaktoren von *S. aureus* und deren Auswirkungen.

Faktor	Auswirkung
Protein A, Matrix-bindende Proteine (MSCRAMM): Clumping Factor A (ClfA), Fibronectin-Bindungsproteine (FnbpA & FnbpB), Kollagen-Bindungsprotein (Can)	Adhäsion an Gewebe und Polymeroberflächen
Toxine: Haemolysine, Leukozidine, PVL (Panton-Valentine Leukozidin), Leukotoxin, Exfoliatine A und B (schwache Superantigene)	Zerstörung der Zellmembran der Wirtszelle
Exoenzyme: DNase, Hyaluronidase, Lipasen, Serinprotease, Kinase	Zellinvasion ohne starke Immunabwehr
Superantigene: TSST-1, Enterotoxine (*sea*, *seb*, *sec1-3*, *sed*, *see*, *seh*)	Auslösen der inflammatorischen Immunanwehr durch unkontrollierten Zytokinausstoß
sak, *chp*, *scn*, *aur*, Protein A, Polysaccharidkapsel, Koagulase, Katalase	Überleben im Wirt/Immunabwehr

1.2 Methicillin-resistente *Staphylococcus aureus*

Wie bereits in Kapitel 1.1 *Krankheitsbilder* erwähnt, geht eine besondere Bedrohung von *S. aureus*-Stämmen aus, die durch Mutationen oder Aufnahme von verschiedenen Determinanten (z.B. der SCC*mec*-Kassette) gegen Antibiotika resistent geworden sind. Schon in den frühen 1960er Jahren, nur wenige Jahre nach der Einführung des β-Lactam-Antibiotikums Methicillin, wurden in britischen (Jevons 1961, Knox 1961) und dänischen (Eriksen & Erichesen 1964) Krankenhäusern erste Methicillin-resistente *S. aureus*-Stämme (MRSA) isoliert. MRSA zeichnen sich durch eine Resistenz gegen alle β-Lactam-Antibiotika aus.
Aufgrund von ähnlichen Makrorestriktionsmustern, Multilokus Sequenztypen (MLSTs) und Antibiogrammen konnten starke Ähnlichkeiten zwischen „Paediatric“ und „New York“ MRSA (veraltete Klassifikation aufgrund von Makrorestriktionsanalysen) sowie MRSA aus Dänemark gefunden werden, die alle zwischen 1957 und 1973 isoliert wurden (Crisóstomo *et al.* 2001). Aufgrund diesen Zusammenhangs und der schon beschriebenen Tatsache, dass Dänemark zu den Ländern zählt, in denen MRSA zuerst auftauchte, wurde lange angenommen, MRSA sei in Dänemark entstanden und habe sich von dort über die ganze Welt ausgebreitet (Crisóstomo *et al.* 2001).
Neueste populationsbiologische Untersuchungen haben dagegen gezeigt, dass MRSA nicht nur einmal entstanden sind, sondern dass es eine mehrfache und voneinander unabhängige Entwicklung gab (Nübel *et al.* 2008).

Typen von MRSA. MRSA lassen sich - aus epidemiologischer Sicht - in drei Klassen einteilen:

1. HA-MRSA - „*hospital acquired*“ MRSA, Krankenhaus-assoziiert, Krankenhaus-

spezifische Risikofaktoren

2. CA-MRSA - „*community acquired*“ MRSA (Chambers 2001), ursprünglich unabhängig von Krankenhäusern und mit ihnen assoziierten Risikofaktoren, charakteristisch (wenn auch nicht obligat) ist die Produktion des Toxins PVL (Panton-Valentine Leukozidin), PVL: verantwortlich für nektrotisierende Pneumonie und tiefgehende Hautinfektionen (Gillet *et al.* 2002, Johnsson *et al.* 2004)

3. LA-MRSA - „*livestock associated*“ MRSA - ursprüngliches Reservoir bei Masttieren

Diese Arbeit beschäftigt sich hauptsächlich mit HA-MRSA.

Resistenzmechanismen. Durch den hohen Antibiotikaeinsatz in Krankenhäusern und dem damit verbundenen Selektionsdruck, kommt es zur schnellen Entwicklung von Resistenzen und deren Ausbreitung. Vor allem mobile genetische Elemente (MGEs) spielen hier eine besondere Rolle, da sie durch horizontalen Gentransfer schnell zwischen zwei Stämmen ausgetauscht werden können. Eine besonders wichtige Rolle spielt das MGE „Staphylococcal cassette chromosome *mec*“ (SCC*mec*), da die Aufnahme die Entstehung von MRSA bewirkt.
Zurzeit lassen sich aufgrund ihrer genetischen Zusammensetzung elf verschiedene SCC*mec*-Typen unterscheiden (International Working Group on the Staphylococcal Cassette Chromosome 2011). Die SCC*mec*-Kassette integriert im *S. aureus* Genom am 3'-Ende des Gens *orfX* und setzt sich aus dem *ccr*-Genkomplex (Mobilität der Kassette), dem *mec*-Genkomplex (*mecA*-Gen verantwortlich für Methicillin-Resistenz) sowie den J-Regionen (können zusätzliche Resistenz-Gene enthalten) zusammen.
Das *mecA*-Gen kodiert ein modifiziertes Penicillin-Bindungsprotein (PBP2a). Von *S. aureus* werden insgesamt vier verschiedene PBPs exprimiert, die alle für die Verknüpfung der Zellwandbausteine verantwortlich sind und Zielmoleküle der β-Lactam-Antibiotika darstellen. Kommt es zur Hemmung der Zellwandsynthese durch das β-Lactam-Antibiotikum übernimmt das durch *mecA* codierte PBP2a diese Funktion, da sich dieses durch eine sehr niedrige Affinität gegenüber β-Lactamen auszeichnet (Pinho *et al.* 2001). Durch diesen Resistenzmechanismus kann die Funktion aller β-Lactame, einschließlich der Cephalosporine, Carbapeneme sowie Monobactame umgangen werden (Katayama *et al.* 2000).
Neben der Resistenz gegen Methicillin können im SCC*mec*-Element zusätzliche Gene integriert sein, die Resistenzen gegen weitere β-Lactame sowie andere Antibiotika-Klassen vermitteln. Das häufig integrierte Transposon *Tn554* trägt Gene zur Erythromycin- sowie Spectinomycin-Resistenz auf sich. Der Stamm MRSA252 enthält Gene, die für die Resistenz gegen Bleomycin und Kanamycin verantwortlich sind (Holden *et al.* 2004).

Therapie. Antibiotika der ersten Wahl zur Behandlung von MRSA-Infektionen sind je nach Antibiogramm Oxazolidine (Linezolid), Glykopeptide (Vancomycin, Teicoplanin), Streptogramine (Quinupristin/Dalfopristin), zyklische Lipopeptide (Daptomycin), Tetracycline sowie Tetracyclin-Analoga (Minocyclin, Tigecyclin).

Epidemiologie. MRSA ist ein Hauptverursacher von nosokomialen Infektionen weltweit und in der Europäischen Union gehen 5 % aller Infektionen auf Krankenhaus-assoziierte MRSA zurück (ECDC: `http://www.ecdc.europa.eu/en/healthtopics/Healthcare-associated_infections/Pages/index.aspx`). Allerdings sind europaweite Unterschiede zu beobachten. Der Anteil von MRSA im Krankenhaus kann zwischen weniger als bzw. gleich 1 % (Island, Niederlande, Skandinavien) und bis zu > 50 % (Malta, Portugal) variieren (EARS-Net Datenbank: Stand 13.10.2011). In Deutschland sind 20 % aller isolierten *S. aureus* Methicillin-resistent (EARS-Net Datenbank: Stand 13.10.2011), wobei auch landesweit starke lokale Unterschiede zu beobachten sind. Im europäischen Vergleich nimmt Deutschland damit einen mittleren Platz ein (Abbildung 1.1).

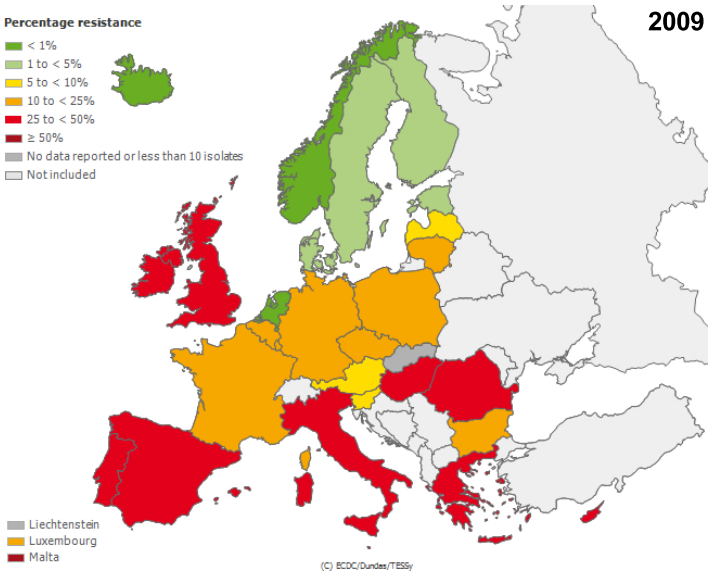

Abbildung 1.1: MRSA-Anteil in europäischen Krankenhäusern im Jahr 2009. (Quelle: EARS-Net Datenbank: Stand 13.10.2011)

In Deutschland ist nach wie vor der Sequenztyp ST22 („Barnim-Epidemiestamm“) am weitesten verbreitet, gefolgt vom Sequenztyp ST225 („Rhein-Hessen-Epidemiestamm“). Beide Stämme sind geographisch gleichmäßig über Deutschland verteilt. War das Vorkommen des Sequenztyps ST45 („Berliner-Epidemiestamm“) in den letzten Jahren stark gesunken, tritt er seit dem Jahr 2010 wieder vermehrt auf (EpiBul 26/2011).
Sowohl weltweit als auch in einzelnen Ländern lässt sich eine Häufung bestimmter epidemischer HA-MRSA-Linien nachweisen. Witte *et al.* (2008) haben gezeigt, dass

Dynamiken existieren, die das Auftreten, die Verbreitung und das Verschwinden verschiedener MRSA-Sequenztypen betreffen (Abbildung 1.2).

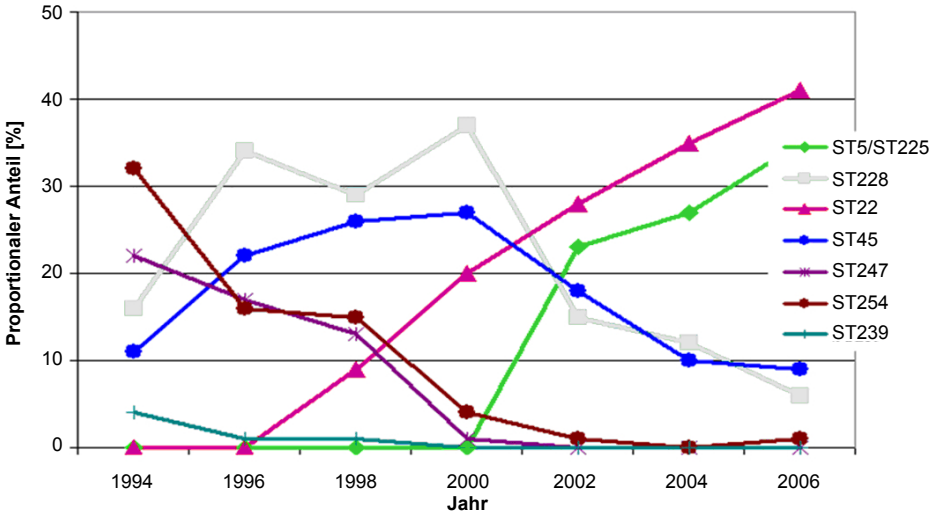

Abbildung 1.2: Dynamiken von in Krankenhäusern auftretenden epidemischen MRSA. (Quelle: Witte *et al.* 2008)

Die genauen Gründe für das temporäre Auftreten sind bisher nicht aufgeklärt, auch wenn dieses Phänomen für viele klonale Mikroorganismen beobachtet wird (Smith *et al.* 1993). Bekannt ist allerdings, dass bereits eine Veränderung der Anzahl vorhandener Klone die Dynamik einer epidemischen Linie ändern können (Cooper *et al.* 2004). Ausgelöst werden können diese Änderungen z.B. durch Evolution der Bakterien, inter- und intraspezifische Konkurrenz (Iwase *et al.* 2010, van Gils *et al.* 2011), Selektion (Gupta *et al.* 1998), Maßnahmen zur Infektionskontrolle (z.B. Isolierung von Patienten, Cooper *et al.* 2004) und den Einsatz von antimikrobiellen Substanzen (Edgeworth 2011).

1.3 Populationsbiologie von *S. aureus* bzw. MRSA

Wie schon in Kapitel 1.2 *Epidemiologie* beschrieben, gibt es HA-MRSA-Linien, die die Population national und global dominieren. Über deren Ursprung und ihre evolutionären Beziehungen war lange Zeit nichts bekannt. Die Erhöhung sowohl von Morbidität und Mortalität als auch die finanzielle Belastung der Gesundheitssysteme und Krankenhäuser, hat in den letzten Jahres ein starkes Interesse an populationsbiologischen Fragestellungen für MRSA aufkommen lassen (Nübel *et al.* 2008).
Die Populationsstruktur von *S. aureus* ist stark klonal aufgebaut. Enright *et al.* (2002) konnten mittels Multilokus Sequenztypisierung (MLST) zeigen, dass die meisten *S. aureus* Sequenztypen und die damit verbundene Diversität in einer limitierten Cluster-Anzahl, den klonalen Komplexen, vorliegen (Abbildung 1.3A). Weiterhin konnte in anderen Publikationen gezeigt werden, dass seit der Einführung von Methicillin im

Jahr 1959 lediglich fünf klonale Komplexe (CC5, CC8, CC22, CC45 und CC30) für den größten Teil nosokomialer Infektionen weltweit verantwortlich sind (Crisóstomo *et al.* 2001, Gomes, Westh & de Lencastre 2006, Conceicao *et al.* 2007, Aires des Sousa *et al.* 2008). Abbildung 1.3B zeigt diese fünf klonalen Komplexe im Detail.

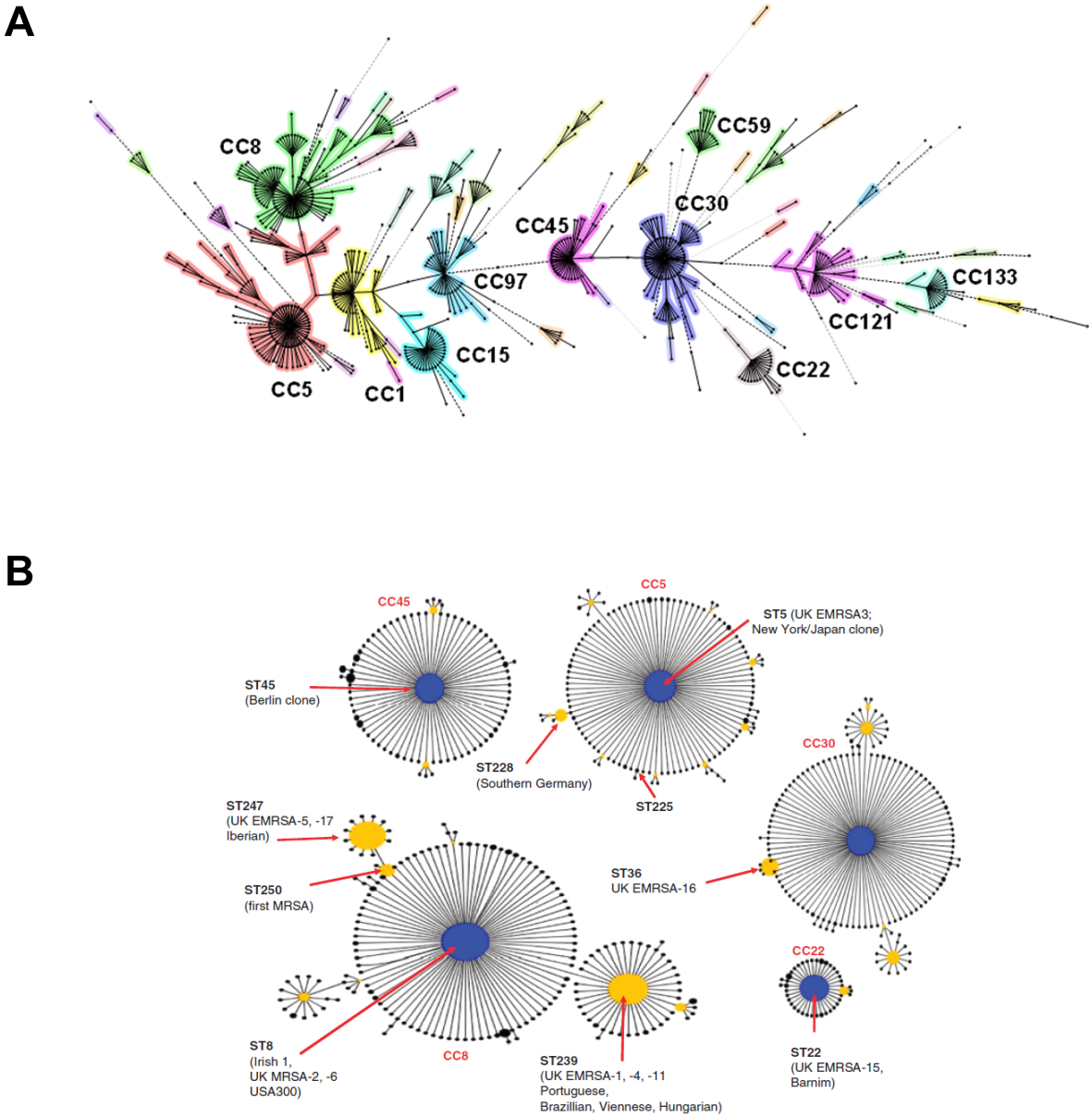

Abbildung 1.3: Populationsstruktur von *S. aureus* basierend auf MLST-Daten. A: Allgemeine Populationsstruktur (Quelle: übernommen von U. Nübel); B: Klonale Komplexe, die für einen Großteil der HA-MRSA-Infektionen verantwortlich sind (Quelle: Willems *et al.* 2011)

Diese Dominanz wirft die Frage auf, was diese klonalen Komplexe so erfolgreich macht. Verschiedene auf MLST beruhende Analysen erklärten das Auftreten von dominanten klonalen Linien innerhalb der globalen MRSA-Population mit der seltenen Aufnahme von SCC*mec* und einer anschließenden Ausbreitung dieser Klone aufgrund von Selektionsvorteilen (Crisóstomo *et al.* 2001, Hiramatsu *et al.* 2001).

Die Entdeckung verschiedener SCC*mec*-Typen sowie die Tatsache, dass Rekombination zwischen SCC*mec*-Elementen selten ist (Lina *et al.* 2006) und Strukturvarianten stabil erhalten bleiben (Chongtrakool *et al.* 2006), führte jedoch zu einem anderen Ergebnis. So konnte durch Vergleiche der Diversität von SCC*mec* und MLST-Analysen gezeigt werden, dass MRSA mehrfach durch die Aufnahme von verschiedenen SCC*mec*-Typen durch unterschiedliche Methicillin-empfindliche *S. aureus* (MSSA)-Linien entstanden sind (Robinson & Enright 2003, Chongtrakool *et al.* 2006, Lina *et al.* 2006).
Diese konträren Ergebnisse zeigen, dass MLST aufgrund der Homogenität der für MLST verwendeten sieben Haushaltsgene von *S. aureus* nicht diskriminierend genug ist, um tiefgehende Fragen zur MRSA-Evolution zu beantworten. Sowohl Holden *et al.* (2004) als auch Witte *et al.* (2008) haben gezeigt, dass gleiche STs verschiedene *spa*-Typen, einen unterschiedlichen Gengehalt als auch medizinisch relevante, phänotypische Unterschiede hinsichtlich ihrer Antibiotika-Resistenz aufweisen können.
Nübel *et al.* (2008) wählten zur Beantwortung der Frage, ob der Erwerb von SCC*mec*-Elementen geographisch und zeitlich unabhängig erfolgt und mit welcher Häufigkeit SCC*mec* aufgenommen wird, einen Ansatz mit höherer Diskriminierung.
Dieser Ansatz beruhte auf der Analyse von 108 Loki (48 kb des *S. aureus* Genoms) und darin enthaltenen Einzelnukleotidpolymorphismen (engl. „*single nucleotide polymorphisms*“, SNPs), die entlang des Genoms durch Punktmutation entstehen. Für den untersuchten Sequenztyp ST5 konnte so gezeigt werden, dass die Anzahl der SCC*mec*-Importe in MSSA mindestens eine Zehnerpotenz höher ist als von Robinson & Enright (2003) angenommen. Die geographische Ausbreitung von MRSA über weite Entfernungen und über kulturelle Grenzen hinweg, ist außerdem, verglichen mit der hohen Frequenz der Aufnahme von SCC*mec*, ein eher seltenes Ereignis (Nübel *et al.* 2008).

SNP-basierte Analysen zur Aufklärung phylogenetischer Zusammenhänge wurden auch schon für andere Pathogene wie z.B. *Yersinia pestis* (Achtman *et al.* 2004), *Mykobacterium tuberculosis* (Alland *et al.* 2003, Baker *et al.* 2004), *Bacillus anthracis* (Pearson *et al.* 2004) und *Salmonella enterica* Typhi (Roumagnac *et al.* 2006) angewendet. Durch die rasanten Fortschritte in der Entwicklung neuer Hochdurchsatz-Sequenziertechnologien, ist es seit kurzem möglich Genom-weite SNPs bzw. ganze Genomsequenzen zu verwenden. Somit ist es möglich, populationsbiologische sowie epidemiologische Fragestellungen mit einer größtmöglichen Auflösung zu beantworten. Erste Genom-basierte Untersuchungen wurden bereits für *Salmonella enterica* Typhi (Holt *et al.* 2008) und *Streptococcus pyogenes* (Beres *et al.* 2009) durchgeführt.

Harris *et al.* (2010) nutzten einen Hochdurchsatz-Genomik-Ansatz, um die Epidemiologie und Mikroevolution des epidemischen MRSA Sequenztyps ST239 aufzuklären, der sich seit Anfang der 60er Jahre des 20. Jahrhundert global ausbreitet. Der ver-

wendete Datensatz setzte sich dabei aus zwei Gruppen zusammen: Die erste Gruppe umfasste einen geographisch und zeitlich distinkten Datensatz, um die globale Populationsstruktur des Sequenztyps ST239 untersuchen zu können; die zweite Gruppe beinhaltete Isolate aus einem thailändischen Krankenhaus, um mögliche Übertragungswege aufzuklären. Der Ursprung des Sequenztyps ST239 ist wahrscheinlich Europa, von wo aus er sich nach Südamerika und Thailand ausgebreitet hat. Aber auch gelegentliche Reimporte nach Europa konnten gezeigt werden. Weniger als 1 % der SNPs, die zur Rekonstruktion der Phylogenie verwendet wurden, waren homoplastisch und ein Viertel dieser SNPs traten in Genen auf, die eine Rolle in der Ausbildung von Antibiotikaresistenzen spielen. Die Daten bestätigen damit frühere Vermutungen, dass der Einsatz von Antibiotika die Evolution von Pathogenen vorantreibt (Aldeyab *et al.* 2008). Der Vorteil des Einsatzes von „*Next Generation Sequencing*"-Technologien ist damit offensichtlich, da eine ähnlich hohe Auflösung mit zur Zeit angewendeten Typisierungsmethoden nicht erreicht werden kann.
Neben der Verwendung von genomweiten SNPs zur Rekonstruktion von evolutionären Zusammenhängen innerhalb epidemischer Linien, ist auch die Kenntnis über die genetische Ausstattung enorm wichtig. Sie liefert weitere Hinweise, um die Epidemiologie von pathogenen Bakterien zu verstehen und nachverfolgen zu können. Das Wissen, welche Veränderungen im Genom zur Ausbildung von Resistenzen führen oder welche Determinanten einem Klon Vorteile bringen, sind also auch ein erster Schritt um neue Strategien in der Bekämpfung von MRSA entwickeln zu können. Kapitel 1.5 gibt eine kurze Einführung in die Genomik von MRSA.

1.4 Der klonale Komplex CC5

Der klonale Komplex CC5 ist einer der fünf klonalen Komplexe, die den Großteil der epidemischen HA-MRSA-Klone umfassen und ist somit weltweit für die Mehrheit der mit MRSA assoziierten Infektionen in Krankenhäusern verantwortlich. Er setzt sich aus den Sequenztypen ST5, ST225 und ST228 sowie vielen Einzel-Lokus Varianten (engl. „*Single Locus Variant*", SLV) zusammen, wobei der Sequenztyp ST5 die Gründerpopulation darstellt.
Der Sequenztyp ST228 wurde anhand seines *Sma*I-Makrorestriktionsmusters seit 1993 in Deutschland beobachtet (Witte *et al.* 1994), später aber auch in Slowenien, Österreich und Italien isoliert (Deurenberg *et al.* 2007, Krziwanek *et al.* 2008). In Italien ist der Sequenztyp ST228 der häufigste HA-MRSA-Klon (Campanile *et al.* 2009), wogegen er sich in Spanien gerade anfängt auszubreiten (Mick *et al.* 2010). Bis zum Jahr 2000 war der Sequenztyp ST228 auch in Deutschland der vom „Nationalen Referenzzentrum für Staphylokokken" (NRZ Staphylokkoken) am häufigsten isolierte Sequenztyp; heute liegt der Anteil gerade noch bei 4 %. Das plötzliche Verschwinden des Sequenztyps

ST228 steht im Zusammenhang mit dem Auftauchen des Sequenztyps ST225, der im Jahr 2006 schon einen Anteil von ca. 35 % der Isolate im NRZ Staphylokokken ausmachte (Abbildung 1.2 in Kapitel 1.2 *Epidemiologie*, Witte *et al.* 2008). Im Jahr 2010 war der Sequenztyp ST225 mit 59 % der zweithäufigste Epidemiestamm in deutschen Krankenhäusern (EpiBul 26/2011). Im Gegensatz zu Deutschland (6 %, EpiBul 26/2011) ist ST5 in den USA, Argentinien und Kolumbien prävalent (Tenover & Goering 2009, Chambers & DeLeo 2009).
Die Häufigkeit des Auftretens des klonalen Komplexes CC5 bringt ein großes Interesse an der Ausbreitung, Populationsstruktur und Evolution dieses klonalen Komplexes mit sich. In den letzten Jahren beschäftigten sich zwei Publikationen näher mit genau diesem Thema. Nübel *et al.* 2008 konnten z.B. zeigen, dass die geographische Ausbreitung von Haplotypen in ST5 phylogeographische Zusammenhänge hat und die Populationsstruktur stark lokal ist (Abbildung 1.4A). In einer weiteren Publikation rekonstruierten Nübel *et al.* 2010 die Populationsstruktur des Sequenztyps ST225, die - im Gegensatz zu dem Sequenztypen ST5 - wenig Diversität aufweist (Abbildung 1.4B). Die geringe Diversität des Sequenztyps ST225 bestätigt, dass der Sequenztyp ST5 die anzestrale Population ist. Die große Nähe zu US-amerikanischen Isolaten, Analysen zur räumlichen Ausbreitung des Sequenztyps ST225 sowie die geringere Diversität unter europäischen Isolaten im Gegensatz zur Diversität in amerikanischen Isolaten haben einen US-amerikanischen Ursprung aufgedeckt.
In dieser Arbeit sollen, aufgrund der Relevanz dieses Klons, weitere Fragen zur Mikroevolution des klonalen Komplexes CC5 und des Sequenztyps ST225 beantwortet werden.

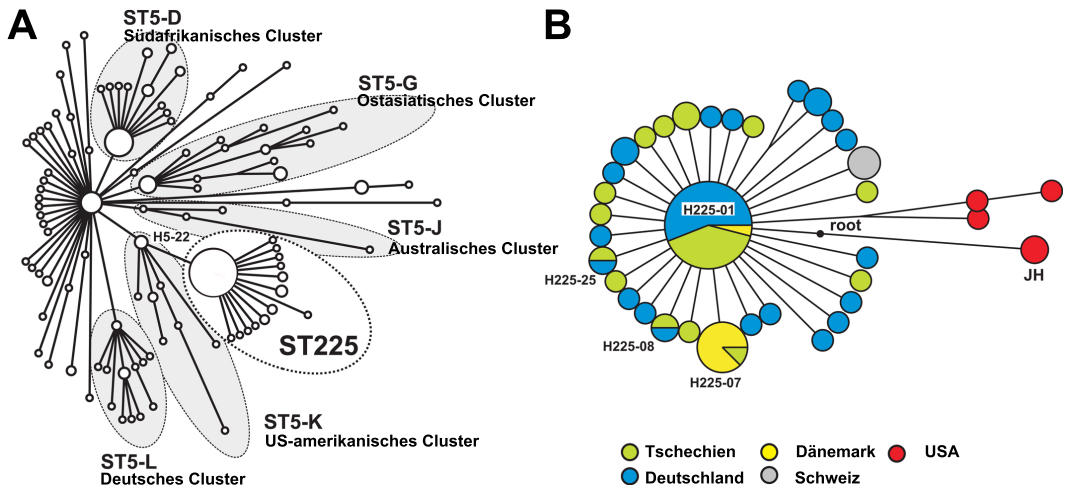

Abbildung 1.4: Populationsstruktur der (A) Sequenztypen ST5 und ST225 und (B) des Sequenztyps ST225 im Detail. (Quelle: Nübel *et al.* 2010)

1.5 Genomik von *S. aureus*

Im Jahr 1995 wurde mit *Haemophilus influenzae* das erste bakterielle Genom sequenziert (Fleischmann *et al.* 1995). 16 Jahre später sind bereits 1.674 komplette sowie weitere 5.140 unvollständige bakterielle Genome in GenBank zu finden (Stand: November 2011).
Vor allem die Entwicklung von neuen Sequenziertechnologien hat zu dieser rasanten Entwicklung beigetragen (Schuster 2008) und schon bald ist es möglich eine „Genomische Enzyklopädie der Bakterien und Archaea“ anzufertigen (Wu *et al.* 2009).
Als einer der wichtigsten Auslöser nosokomialer Infektionen bestand schon früh ein großes Interesse an der kompletten genetischen Ausstattung von *S. aureus*. Bereits 2001 veröffentlichten Kuroda *et al.* die Genome der japanischen Isolate N315 und Mu50. Mittlerweile beinhaltet GenBank Daten für 28 komplette und 68 ungeschlossene Genome sowie Sequenzierrohdaten für weitere 98 Genome. Tabelle 1.2 enthält eine Übersicht der 23 vollständigen und bereits veröffentlichten Genome.

Tabelle 1.2: Details zu den 23 sequenzierten *S. aureus* Genomen.

Stamm	Klonaler Komplex	Sequenz-typ	Wirt	Herkunfts-land	Jahr	Resistenztyp*	Akzessions-nummer
MSSA476	1	1	Mensch	UK	1998	CA-MSSA	BX571857
MW2	1	1	Mensch	USA	1998	CA-MRSA	BA000033
ED98	5	5	Huhn	UK	1996-1997	LA-MRSA	CP001781
Mu3	5	5	Mensch	Japan	1996	heteroVISA	AP009324
Mu50	5	5	Mensch	Japan	1997	HA-VISA	BA000017
N315	5	5	Mensch	Japan	1982	HA-VSSA	BA000018
JH1	5	105	Mensch	USA	2000	HA-VISA	CP000736
JH9	5	105	Mensch	USA	2000	HA-VRSA	CP000703
04-02981	5	225	Mensch	Deutschland	2004	HA-MRSA	CP001844
NCTC8325	8	8	Mensch	UK	<1949	Laborstamm	AP009351
USA300_FPR3757	8	8	Mensch	USA	2000 (?)	CA-MRSA	CP000255
USA300_TCH1516	8	8	Mensch	USA	2002-2004	CA-MRSA	CP000730
COL	8	205	Mensch	UK	1961	früher MRSA	CP000046
T0131	8	239	Mensch	China	n.a.	MRSA	CP002643
JKD6008	8	239	Mensch	Neuseeland	2003	VISA	CP002120
TW20	8	239	Mensch	UK	2003	HA-MRSA	FN433596
Newman	8	254	Mensch	UK (?)	1952	n.a.	AP009351
MRSA252	30	36	Mensch	UK	1997	HA-MRSA	BX571856
MSHR1132	75	1850	Mensch	Australien	2006	CA-MRSA	FR821777
ED133	133	133	Schaf	Frankreich	1997	n.a.	CP001996
ET3-1/RF122	151	151	Rind	Irland	n.a.	n.a.	AJ938182
SO385	398	398	Mensch	Niederlande	2006	LA-MRSA	AM990992
JKD6159	-	93-IV	Mensch	Australien	n.a.	CA-MRSA	CP002114

n.a. - Information nicht verfügbar. *Abkürzungen Resistenztyp: MRSA - Methicillin-resistente *S. aureus*; MSSA - Methicillin-sensitive *S. aureus*; VISA - Vancomycin-intermediär-sensible *S. aureus* ; VSSA - Vancomycin-sensitive *S. aureus*; VRSA - Vancomycin-resistente *S. aureus*

Die 23 Genome variieren in ihrer Größe zwischen 2,74 und 3,04 Mb und enthalten zwischen 2.555 und 2.892 Gene. Die Mehrheit der Gene weist ein hohes Maß an DNA-Identität auf (> 97 %). Paarweise Vergleiche verschiedener *S. aureus*-Stämme haben einen kolinearen Aufbau des zirkulären Genoms gezeigt, der lediglich durch die Integration von einigen unterschiedlichen Elementen unterbrochen wird. Aus Abbildung 1.5 ist klar erkenntlich, dass das *S. aureus*-Genom aus zwei großen Komponenten besteht: dem stabilen Kerngenom, dessen Gene in allen Stämmen vorkommen, sowie dem variablen oder akzessorischen Genom (Lindsay & Holden 2004).

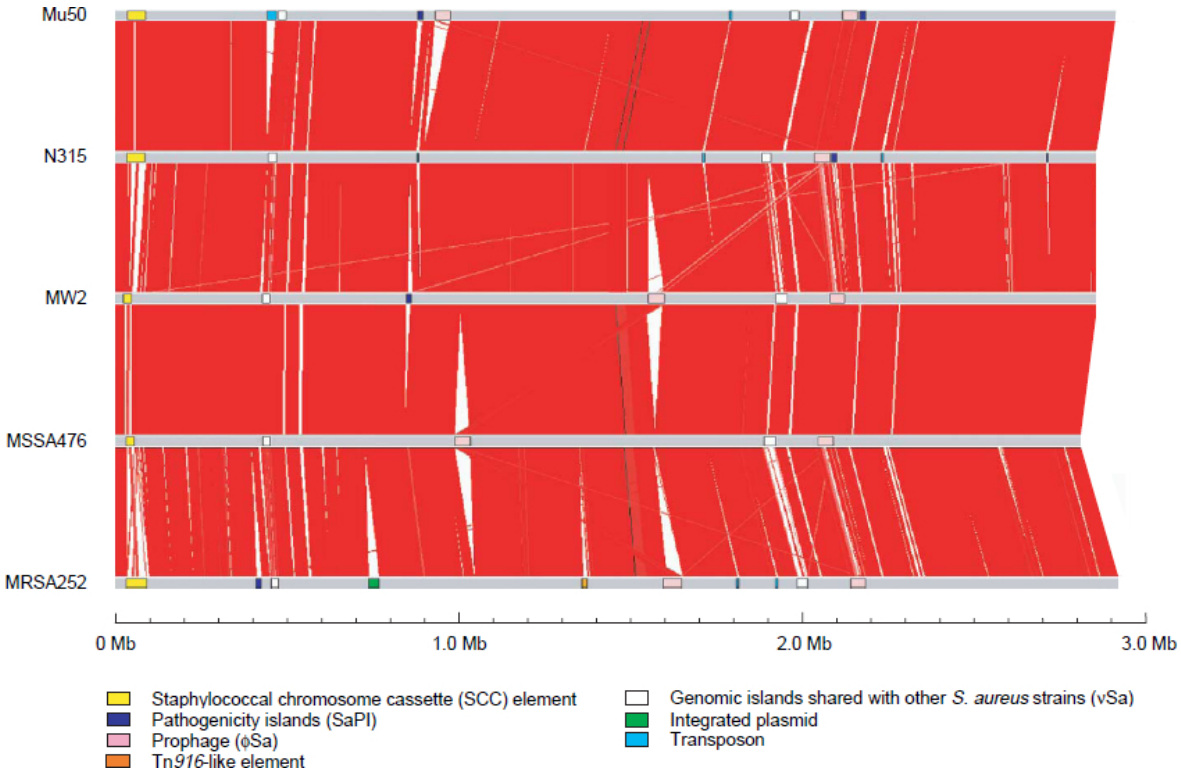

Abbildung 1.5: Vergleich von fünf *S. aureus*-Genomen. (Quelle: Lindsay & Holden 2004)

1.5.1 Das Kerngenom

Das Kerngenom macht etwa 75 % des kompletten *S. aureus* Genoms aus (Lindsay & Holden 2004) und setzt sich vor allem aus Genen des zentralen Metabolismus und weiteren Haushaltsgenen zusammen. Aber auch Gene, die Oberflächenbinde-Proteine, Toxine, Exoenzyme sowie Virulenzfaktoren kodieren, die nicht in anderen Staphylokkoken-Spezies vorkommen, gehören dem Kerngenom an. Trotz der starken Konservierung des Kerngenoms, sollten kleine Veränderungen nicht vernachlässigt werden, da sie starke Auswirkungen auf die Genexpression und die Proteinfunktion haben können. Genetische Diversität im Kerngenom kann somit für phänotypische Unterschiede zwischen den Stämmen verantwortlich sein (Lindsay & Holden 2006).
Sequenzunterschiede im Kerngenom entstehen vor allem durch Einzelnukleotidpolymorphismen (SNPs) und die Anzahl von diversifizierenden SNPs hängt von der Distanz der untersuchten Isolate ab (Tabelle 1.3).

1.5.2 Das akzessorische Genom

Das akzessorische Genom besteht vor allem aus Genen, die keine essentielle Funktion haben. Diese umfasst Virulenz, Medikamenten- und Metallresistenz, Substratverwertung und diverse Metabolismen. Viele Bereiche, die das variable Genom ausmachen, sind mobile genetische Elemente, die horizontal übertragen werden und somit zu ei-

Tabelle 1.3: SNPs im Kerngenom von *S. aureus*.

Vergleich	SNPs im Kerngenom	SNPs in kodierenden Regionen
innerhalb eines klonalen Komplexes		
N315 (ST5, CC5) : Mu50 (ST5, CC5)	315	223
N315 (ST5, CC5) : JH1 (ST105, CC5)	508	386
zwischen verschiedenen klonalen Komplexen		
N315 (ST5, CC5) : MW2 (ST1, CC1)	20.252	16.758
N315 (ST5, CC5) : USA300 (ST8, CC8)	20.262	16.765

nem Großteil der Diversität zwischen den einzelnen Stämmen beitragen. Diese Elemente umfassen Prophagen, Pathogenitätsinseln, Chromosomale Kassetten (SCC*mec*), Transposons und Plasmide. Viele dieser Elemente tragen Pathogenitäts-, Virulenz- oder Resistenzfaktoren. Interessanterweise sind Virulenzgene mit Prophagen und Pathogenitätsinseln assoziiert, wogegen Resistenzgene eher auf der SCC*mec*-Kassette, Plasmiden und Transposons zu finden sind (Lindsay & Holden 2006).
Die Identifikation und Charakterisierung von mobilen genetischen Elementen sowie Erkenntnisse über ihre Verbreitung in der Population können helfen, ihre Evolution sowie ihre Rolle in der Pathogenese von *S. aureus* zu verstehen. Dieses Wissen hat somit eine direkte klinische Implikation (Lindsay & Holden 2004). Da *S. aureus* nicht in der Lage ist, Gene via Transformation zu übertragen, und Konjugation nur in sehr geringen Umfang stattfindet, ist der Großteil des horizontalen Transfers wahrscheinlich von der Transduktion durch Bakteriophagen abhängig (Lindsay & Holden 2006, Lindsay 2010). Bakteriophagen bzw. Prophagen spielen somit vermutlich eine entscheidende Rolle in der Evolution von *S. aureus* (Lindsay & Holden 2006).

1.6 Prophagen

Bakterien waren lange die einzigen Lebewesen auf der Erde und ihre Evolution umfasste daher nur intra- und interspezifische Konkurrenz, genetischen Austausch und Selektion. Da der evolutionäre Ursprung von Phagen wahrscheinlich nicht weit von dem ihres bakteriellen Wirts entfernt ist, kann aber angenommen werden, dass auch Phagen an der Evolution von Bakterien beteiligt waren (Brüssow, Canchaya & Hardt 2004, Hatfull 2008).
Bakteriophagen stellen - mit einer Populationsgröße von 10^{31} - das größte Reservoir unbekannter genetischer Informationen dar (Wommack & Colwell 2000). Berechnungen haben ergeben, dass alle Bakterien der Erde zusammen von etwa 10^{25} viralen Infektionen pro Sekunde betroffen sind (Pedulla *et al.* 2003). Somit wird die gesamte Phagen-Population alle paar Tage komplett ausgetauscht. Die Population ist dadurch extrem dynamisch und jeder Infektionszyklus besitzt das Potential, veränderte oder mutierte Phagen zu erzeugen (Pedulla *et al.* 2003, Hatfull 2010).

Durch ihre Eigenschaft neue biochemische und physiologische Merkmale in die Bakterien einzubringen, können Phagen den Phänotyp oder die Fitness des Wirts verändern. Es ist außerdem schon lange bekannt, dass Phagen Gene besitzen, die für ihren Lebenszyklus nicht gebraucht werden, aber in Pathogenen als Virulenzfaktoren große Bedeutung haben (Freeman 1951, Barksdale & Arden 1974).
Als Konsequenz daraus wird angenommen, dass Phagen eine große Rolle in der Adaption von Pathogenen an neue Wirte und der Entstehung neuer Pathogene oder epidemischer Klone spielen (Wagner & Waldor 2002, Brüssow, Canchaya & Hardt 2004).
Auch in *S. aureus* kommen Phagen sehr häufig vor. Als mobiles genetisches Element sind sie dabei als Prophage in das Wirtsgenom integriert. Da sich ein Teil dieser Arbeit mit der Analyse von Mechanismen der Prophagenevolution beschäftigt, gibt dieses Kapitel einen kurzen Überblick in die Thematik.

1.6.1 Taxonomie

Phagen in *S. aureus* gehören der Ordnung der Caudovirales an, die durch ein Helixförmiges Schwanzstück sowie einen Ikosaeder-förmigen Kopf charakterisiert ist. Die Caudovirales setzt sich aus den drei Familien *Myoviridae* (langes kontraktiles Schwanzstück), *Siphoviridae* (langes nicht-kontraktiles Schwanzstück) und *Podoviridae* (kurzes nicht-kontraktiles Schwanzstück) zusammen, wobei die *Siphoviridae* den größten Anteil in *S. aureus* ausmachen (Kwan *et al.* 2005).

1.6.2 Das Phagen-Genom

Das Genom der *Siphoviridae* in *S. aureus* ist in acht Modulen organisiert, die alle eine spezifische Aufgabe übernehmen: Lysogenie, DNA-Replikation, Regulation der Transkription, DNA-Verpackung und Kopf, Synthese der Kopf-Schwanz-Verbindung, Synthese des Schwanzes, Synthese der kurzen Schwanzfasern und Wirts-Lyse (Brüssow & Desiere 2001).

Abbildung 1.6: Modularer Aufbau des *S. aureus* Prophagen ϕN315. (Quelle: Brüssow, Canchaya & Hardt 2004)

Unterschiede in Phagen-Genomen sind meist durch den Austausch von DNA zu erklären (Brüssow, Canchaya & Hardt 2004) und der bereits beschriebene modulare Aufbau des Phagen-Genoms erleichtert dieses. Ein Modul kann einfach durch ein Sequenzunterschiedliches Modul eines anderen Phagen ersetzt werden, solange dieses die gleiche Funktion besitzt. Die Module gleicher Funktion zeichnen sich in *S. aureus* durch ein

hohes Maß einer mosaikartigen Struktur aus (Kwan *et al.* 2005). Regionen mit offensichtlicher Sequenzidentität sind dabei immer wieder von nicht-verwandten Sequenzregionen unterbrochen, was auf einen beträchtlichen horizontalen genetischen Austausch hindeutet (Hendrix *et al.* 1999). Dies erklärt die große genetische Diversität in den Phagen (Canchaya *et al.* 2003).

Rolle der Prophagen für die Pathogenität von *S. aureus*. Einige in *S. aureus* vorkommende Prophagen besitzen Gene, die Toxine kodieren (Iandolo *et al.* 2002, Brüssow, Canchaya & Hardt 2004, Pantucek *et al.* 2004) und somit eine große Rolle in der Pathogenität spielen.
Diese zusätzlichen Gene sind oft in der Nähe der rechten Anheftungsstelle (*attR*) des Phagen-Genoms zu finden. Es wird angenommen, dass solche Gene zufällig durch fehlerhaftes Ausschneiden des Prophagens aus dem bakteriellen Genom im Phagen bleiben und sich so verbreiten (Ferretti *et al.* 2001).
Neben dieser positiven lysogenen Konversion kann die Integration eines Prophagen aber auch eine negative Auswirkung auf die Pathogenität des Wirts haben. So kommen in *S. aureus* Prophagen vor, die die Gene des Invasins Lipase (*geh*) und des Toxins β-Hämolysin (*hlb*) durch Insertion zerstören und so die Pathogenität herabsetzen können (Coleman *et al.* 1991).

1.6.3 Evolution

Untersuchungen zur Evolution von Phagen sind sehr schwierig und die bisherigen Ergebnisse variieren stark. Gründe hierfür sind die große Phagenpopulation bei einer nur geringen Zahl bisher sequenzierter Phagen-Genome sowie die extrem schnelle Veränderung der Genome (Brüssow, Canchaya & Hardt 2004). Es wird allerdings angenommen, dass vor allem horizontaler DNA-Austausch für die Diversifizierung von Phagen verantwortlich ist.

Zeit/Datierung. Bisher konnte der Zeitpunkt, wann horizontale Evolution stattfindet, nicht aufgeklärt werden. Zwei Möglichkeiten sind vorstellbar. Nach der ersten Hypothese ist der Großteil der aktuell zu beobachtenden Diversität schon sehr früh in der Geschichte der Phagen entstanden. Ab einem bestimmten Punkt in der evolutionären Geschichte - z.B. dem Beginn der bakteriellen Artbildung - wurde sie dann durch vertikale Transmission an heutige Phagen weitergegeben. Bei der Annahme von einem höchst diversen Phagen-Genpool, kann die Ähnlichkeit von Genen zwischen einigen Phagen trotz großer phylogenetischer Wirts-Distanzen einfach mit der großen Anzahl verschiedener Genkombinationen zu dieser Zeit erklärt werden. Eine andere - wahrscheinlichere - Hypothese ist, dass horizontaler Austausch bis heute andauert. Die Erhaltung der Sequenzähnlichkeit zwischen einigen Genen über einen langen Zeitraum

hinweg ist sehr unwahrscheinlich, wenn man die starke Divergenz zwischen anderen Genen betrachtet (Hendrix *et al.* 1999).

Mechanismen der Diversifizierung. Die genetischen Mechanismen, die für den Austausch von Genen oder Modulen zwischen Phagen und damit für ihre Evolution, zuständig sind, sind bisher nicht aufgeklärt. Diskutiert werden drei Rekombinationstypen, die in Abbildung 1.7 dargestellt sind.

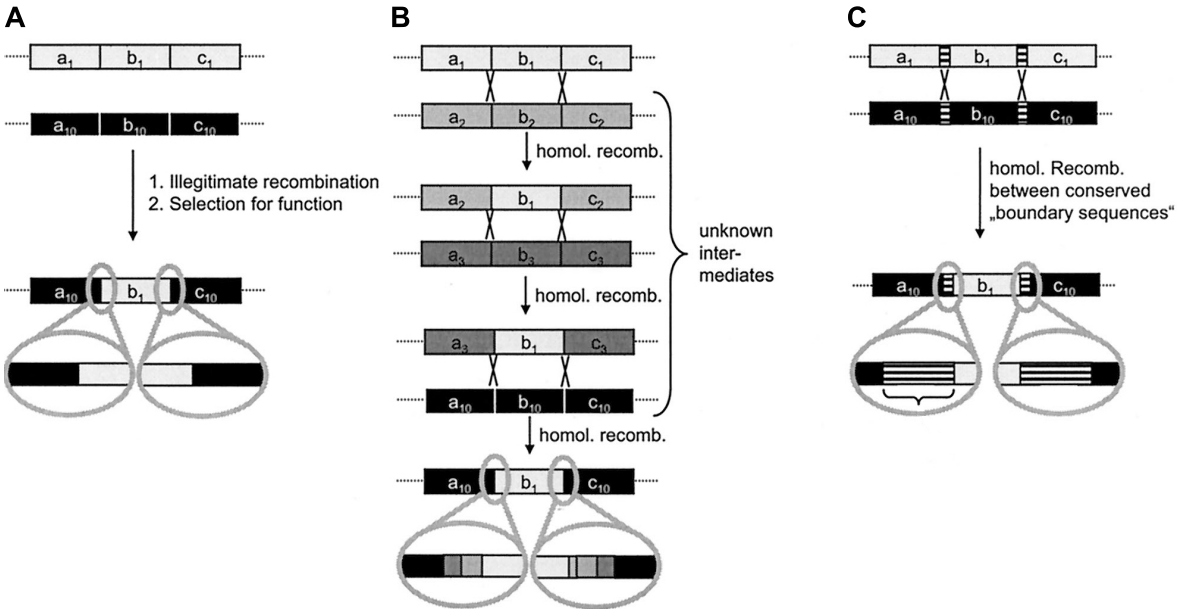

Abbildung 1.7: Molekulare Modelle für den Austausch von DNA-Abschnitten zwischen Phagen. A: nicht-homologe Rekombination; B: sich wiederholende homologe Rekombination; C: sequenzspezifiscche Rekombination (Quelle: Brüssow, Canchaya & Hardt 2004)

Nicht-homologe Rekombination tritt wahllos und überall im Genom auf. Dies führt dazu, dass Rekombination innerhalb von Genen auftritt, das Phagengenom zu stark vergrößert wird oder Gencluster zerstört werden. Die folgende stringente Selektion auf funktionsfähige Phagen (Juhala *et al.* 2000) sorgt dafür, dass die meisten Produkte nicht-homologer Rekombination in kodierenden Regionen entfernt werden und nur noch Phagen übrig bleiben, die an den Gengrenzen rekombiniert sind. Dieser Prozess ist damit nicht besonders erfolgreich, da nur wenige „lebensfähige“ Phagen entstehen (Brüssow & Hendrix 2002).

Für λ-förmige Phagen wurde gezeigt, dass sich in ihrem Genom kleine, homologe Sequenzabschnitte an Gengrenzen befinden (Clark *et al.* 2001). Solche Linkersequenzen können den Austausch von DNA durch homologe oder sequenzspezifische Rekombination herbeiführen. Diese beiden Arten von Rekombination können immer dann auftreten, sobald ein Phage einen Wirt infiziert, der einen Prophagen mit ausreichender Homologie in sich trägt. Durch homologe bzw. sequenzspezifische Rekombination können sich Gene bzw. Module (Gengrenzen korrelieren mit Modulgrenzen) somit sehr schnell

in der Population ausbreiten. Aus diesem Grund werden diese Mechanismen als die antreibende Kraft für die Evolution von Phagen angenommen (Brüssow & Hendrix 2002).

Zugang zum Sequenzpool. Die Größe des Sequenzpools, auf den Phagen zugreifen können, ist nach wie vor unbekannt (Hendrix 1999, Goerke *et al.* 2009). Theoretisch haben alle Phagen über horizontalen DNA-Austausch Zugang zu einem riesigen, gemeinsamen Sequenzpool. Allerdings ist die Erreichbarkeit dieses Pools abhängig von der Anzahl der Barrieren zwischen einer Sequenz, dem Phagen und somit der Anzahl individueller Schritte, die zu einem Austausch führen können. So können z.B. enge Wirtsgrenzen die Phagen daran hindern, genetisches Material auszutauschen (Hendrix 1999). Es ist auch möglich, dass sich individuelle Pools entwickelt haben (Goerke *et al.* 2009). Zur Beantwortung dieser Frage, müssen mehr Daten analysiert werden.

1.6.4 Praktische Anwendung

Die Erforschung von Bakteriophagen ist nicht nur aus wissenschaftlicher Sicht interessant. Vielmehr haben Phagen auch einen praktischen Nutzen und werden in der Gentechnik (Phagen als Vektor), der Biotechnologie (Phagen-Display), der Agrartechnologie (Transduktion von Genen in Nutzpflanzen), der mikrobiologischen Diagnostik und der Medizin eingesetzt. Auf die letzten beiden Anwendungsgebiete soll im Folgenden kurz eingegangen werden.

Mikrobiologische Diagnostik. In der Diagnostik werden Phagen zur Bestimmung von pathogenen Bakterien eingesetzt. Dabei macht man sich die Wirtsspezifität der Phagen zu Nutzen. Kultivierte Bakterien werden mit spezifischen Phagen infiziert und auf die Lyse der Bakterien durch den Phagen untersucht. Diese „Lysotopie“ genannte Methode wird heute vor allem für die Unterscheidung verschiedener *Salmonella* Serovare verwendet.

Medizin. Lytische Bakteriophagen können auch zur örtlichen oder systemischen Behandlung von bakteriellen Infektionen verwendet werden (Chibani-Chennoufi *et al.* 2004). Gegenüber Antibiotika haben sie den Vorteil, dass sie spezifischer sind, gegen resistente Pathogene eingesetzt werden können und weder Mensch noch die Normalflora schädigen.
Obwohl der Einsatz von Phagen zu therapeutischen Zwecken in der ehemaligen Sowjetunion und Osteuropa seit vielen Jahren erfolgreich und dort gut erforscht ist (Sulakvelidze, Alavidze & Morris 2001), ist er im Westen nach wie vor für Menschen verboten. Allerdings ist das Thema in den letzten Jahren auch hier in den Fokus getreten und für *Pseudomonas aeruginosa* wurden bereits die Ergebnisse der klinischen Studien der

Phasen 1 und 2 veröffentlicht (Wright *et al.* 2009). Auch für *S. aureus*-Infektionen berichten einige Publikationen von einer erfolgreichen Heilung durch den Einsatz von Phagen (Sulakvelidze, Alavidze & Morris 2001, Markoishvili *et al.* 2002, Jikia *et al.* 2005).

1.7 Zielsetzung

- Etablierung von bioinformatischen Methoden zur Verarbeitung und Auswertung von Daten der „*next generation sequencing*“-Technologien
- Aufklärung der genetischen Populationsstruktur des klonalen Komplexes CC5 und des Sequenztyps ST225
- Vergleichende Genomanalysen zur Aufklärung der genetischen Ausstattung von CC5-Isolaten
- Untersuchungen zu den Mechanismen, die die Diversität von Prophagen erklären
- Rekonstruktion der räumlichen Ausbreitung von MRSA

2 Material

2.1 Verwendete Isolate

Um die Mikroevolution des klonalen Komplexes CC5 und des Sequenztyps ST225 mit einer möglichst großen Auflösung rekonstruieren zu können, werden für die durchzuführenden Analysen ganze Genomsequenzen von insgesamt 91 *Staphylococcus aureus* Isolaten verwendet. Dabei werden für den klonalen Komplex CC5 24 und für den Sequenztyp ST225 67 Isolate verwendet.
Die Isolate des klonalen Komplexes 5 wurden anhand eines „*Minimum Spanning Trees*“ von Nübel *et al.* (2008) so ausgewählt, dass sie repräsentativ für die Populationsstruktur sind (Kapitel 5.2.1, Abbildung 5.2) und um weitere in GenBank vorhandene Stämme ergänzt.
Detaillierte Informationen zu den Stämmen befinden sich in den Tabellen 5.2 (CC5) und 5.7 (ST225).

2.2 Chemikalien, Enzyme und Kits

10 x PCR-Puffer S	PeqLab, Erlangen
100 bp DNA Ladder, extended	Roth, Karlsruhe
BigDye Terminator v3.1 Cycle Sequencing Kit	Applied Biosystems, Darmstadt
DNeasy Blood & Tissue Kit	Qiagen, Hilden
dNTP Mix	Roth, Karlsruhe
Ethidiumbromid	
GelPilot DNA Loading Dye	Qiagen, Hilden
GS FLX Titanium XLR70 Kit	Roche, Mannheim
Hot Taq-DNA-Polymerase	PeqLab, Erlangen
Oligonukleotid-Primer	Invitrogen, Darmstadt
QIAquick PCR Purification Kit	Qiagen, Hilden

2.3 Puffer und Medien

2.3.1 Reagenzien und Medien zur Anzucht

LB-Medium (1 L, pH 7,4)

Trypton	10,0	g
Hefeextrakt	5,0	g
NaCl	10,0	g

2.3.2 Puffer für Elektrophorese

TAE-Puffer (50-fach) (1 L, pH 8,3)

Tris Base	242,0	g
Essigsäure	57,1	mL
EDTA 0,5 M	100,0	mL

TBE-Puffer (5-fach) (1 L, pH 8)

Tris Base	54,0	g
Borsäure	27,5	mL
EDTA 0,5 M	3,74	mL

2.4 Geräte

Elektrophoresekammer	Sub-Cell GT	BioRad, München
Geldokumentation	GelDoc XR	BioRad, München
Kapillarsequenzierer	3130xl Genetic Analyzer	Applied Biosystems, Darmstadt
Pipetten	Reference	Eppendorf, Hamburg
Pyrosequenzierer	454 GS FLX	Roche, Mannheim
Spannungsquelle	PowerPack Basic	BioRad, München
Spectrophotometer	SmartSpec 3000	BioRad, München
Thermocycler	GeneAmp PCR System 9700	Applied Biosystems, Darmstadt
	PTC-200 DNA Engine	MJ Research, St. Bruno, Kanada
	TProfessional Basic	Biometra, Göttingen
Zentrifuge	Centrifuge 5804R	Eppendorf, Hamburg

2.5 Primer

Die Verifizierung von SNPs erfolgt mittels Sanger-Sequenzierung. Die Primer werden von der Firma Invitrogen, Darmstadt synthetisiert und mit ddH_2O auf eine Konzentration von 10 μMol eingestellt.
Die Tabellen A.1 und A.3 im Anhang enthalten die verwendeten Oligonukleotid-Primer.

2.6 Computerressourcen

Tabelle 2.1 enthält eine Übersicht der verwendeten Rechen- und Speicherkapazitäten.

Tabelle 2.1: Verwendete Computerressourcen.

Betriebssystem	CPU	RAM	Speicher
	Desktopcomputer		
Microsoft Windows XP	Intel Core 2 Duo E8500; 3,16 GHz	3,50 GB	1,9 TB
	Virtuelle Maschinen		
SUSE Linux Enterprise Server 11 SP1 (i586)	Six-Core AMD Opteron^{TM} Processor 2435; 2,60 GHz	7,72 GB	24 GB
SUSE Linux Enterprise Server 11 SP1 (x86_64)	Six-Core AMD Opteron^{TM} Processor 2435; 2,60 GHz	7,75 GB	64 GB

2.7 Software

In der vorliegenden Arbeit werden eine Reihe verschiedener Computerprogramme verwendet, um Sequenzdaten zu analysieren und auszuwerten und phylogenetische Analysen durchzuführen. Tabelle 2.2 enthält eine Übersicht dieser Programme.

Tabelle 2.2: Verwendete Software.

Name	Anwendung	Hersteller
Abacas v1.2	Ausrichtung von Contigs anhand einer Referenz	Freeware (Assefa *et al.* 2009)
Artemis	Annotation von Sequenzdaten	Freeware (Rutherford *et al.* 2000)
BEAST v1.6.1	Phylogenetische Analysen	Freeware (Drummond & Rambaut 2007)
BEAUTi v1.6.1	Generieren von BEAST XML Dateien	Freeware (Drummond, Rambaut & Suchard 2010)
BioLayout *Express*^{3D} Update 8	Clustern von Proteinen	Freeware (Theocharidis *et al.* 2009)
BLAST v2.2.23+	Paarweise Sequenzvergleiche	Freeware (Altschul *et al.* 1990)
ClonalFrame v1.2	Detektion von Rekombinationsereignissen	Freeware (Didelot & Falush 2007)
Dendroscope v2.2.2	Visualisierung und Bearbeitung von phylogenetischen Bäumen	Freeware (Huson *et al.* 2007)
DnaSP v5.10.01	Detektierung von homoplastischen SNPs & Berechnung von dN/dS	Freeware (Librado & Rozas 2009)
Kodon v3.6.1	Whole Genome Alignment, Vergleichende Genomanalysen, SNP Analysen	Applied Maths, Sint-Martens-Latem, Belgien
Mauve v2.3.1	Whole Genome Alignment	Freeware (Darling, Mau & Perna 2010)
MAQ v0.6.8	Mapping von Sequenzdaten	Freeware (Li, Ruan & Durbin 2008)
MEGA v4 & v5	Sequenzalignments	Freeware (Tamura *et al.* 2007, Tamura *et al.* 2010)

Tabelle 2.2: Verwendete Software.

Name	Anwendung	Hersteller
MrBayes v3.1.2	Phylogenetische Analysen	Freeware (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003)
Newbler v2.0	*de novo* Assemblierung von Sequenzdaten	Roche, Mannheim, Deutschland
Path-O-Gen v1.3	Detektion des zeitlichen Signals in Sequenzdaten	Freeware (Rambaut 2011)
PAUP* v4.0b10	Phylogenetische Analysen	Sinauer Associates, Sunderland, USA (Wilgenbusch & Swofford 2003)
Primer3 v0.4.0	Entwurf von Oligonukleotid-Primern	Freeware (Rozen & Skaletsky 2000)
SAMtools v0.1.12	Alignmentverarbeitung	Freeware (Li *et al.* 2009)
SigmaPlot v11.0	Datenauswertung	SPSS Inc., Chicago, USA
SOAPdenovo v1.04	*de novo* Assemblierung von Sequenzdaten	Freeware (Li *et al.* 2010)
SSAHA2 v2.5.3	Read Mapping	Freeware (Ning, Cox & Mullikin 2001)
Tracer v1.5	Visualisierung von MCMC Output	Freeware (Rambaut & Drummond 2007)
Treefinder vOktober2008	Phylogenetische Analysen	Freeware (Jobb, von Haeseler & Strimmer 2004)
zt v1.1	Durchführung von Mantel Tests	Freeware (Bonnet & Van de Peer 2002)

3 Molekularbiologische Methoden

Die angewendeten molekularbiologischen Methoden wurden sowohl von Mitarbeitern des Robert Koch-Instituts (RKI) als auch von externen Firmen ausgeführt:

- Anzucht, Stammhaltung, DNA-Präparation, PCR: Annette Weller, Heike Illiger und Mike Henkel (RKI, Wernigerode)
- Sequenzierung: RKI Berlin, 454 Life Sciences, Georg-August-Universität Göttingen, GATC Konstanz (Tabelle 3.1)

Obwohl die angewendeten molekularbiologischen Methoden nicht von mir selbst durchgeführt wurden, werden sie in den nächsten Kapiteln kurz vorgestellt, da sie die Grundlage der nachfolgenden bioinformatischen Auswertung bilden.

3.1 Anzucht von *Staphylococcus aureus* und Stammhaltung

Die Isolate werden aerob in LB-Medium mit 1 % Glycin bei 37 °C über Nacht in Reagenzgläsern im Schüttelinkubator angezüchtet. Die Zellen werden auf Müller-Hinton-Blutagar-Platten (Oxoid, Cambridge, UK) ausgestrichen und bei 37 °C über Nacht inkubiert und für nachfolgende Experimente bei 4 °C gelagert.

3.2 Extraktion chromosomaler DNA

Für die Isolierung von chromosomaler DNA wird das Qiagen DNeasy Blood & Tissue Kit (Qiagen, Hilden) verwendet. Dieses basiert auf einer selektiven Bindung von Nukleinsäuren an Silicagel-Membranen und nutzt deren Eluation unter niederosmolaren Bedingungen aus. Die extrahierte DNA wird in 100 μL TAE-Puffer eluiert und bei -20 °C gelagert.

3.3 Bestimmung der Konzentration und der Reinheit von DNA

Die Bestimmung der DNA-Konzentration erfolgt spektralphotometrisch durch Absorptionsmessung bei 260 nm im SmartSpec 3000 (BioRad, München) unter Verwendung von Quarzküvetten.
Eine Konzentration von 50 μg/mL dsDNA entspricht einer $OD_{260} = 1$. Der Grad der Reinheit von Nukleinsäuren berechnet sich aus dem Verhältnis der Absorptionskoeffizienten A_{260nm} und A_{280nm}. Quotienten zwischen 1,8 und 2 stehen für eine reine DNA-Lösung. Kleinere Werte deuten auf Verschmutzungen mit Proteinen hin, wogegen höhere Werte ein Zeichen für RNA-Verunreinigungen sind (Sambrook, Fritsch & Maniatis 1989).

3.4 Die Polymerase-Kettenreaktion

Die Polymerase-Kettenreaktion (PCR, Saiki *et al.* 1985) orientiert sich an den natürlichen DNA-Replikationsvorgängen in der Zelle. Diese sind Denaturierung der Ausgangs-DNA, Annealing des Primers und Elongation des neu entstandenen DNA-Strangs. Alle drei Schritte werden mehrfach wiederholt, so dass am Ende der gewünschte DNA-Abschnitt in mehreren Millionen Kopien vorliegt.
Der Ansatz für eine PCR mit einem Reaktionsvolumen von 10 μL besteht aus folgenden Komponenten: 0,2 μL Template-DNA, 1 μL 10x PCR-Puffer S (PeqLab), 0,2 μL dNTPs (10 mM), 0,2 μL 3'-Primer (10μmol), 0,2 μL 5'-Primer (10μmol), 0,05 μL *Taq*-Polymerase (5U/μL, PeqLab), 8,15 μL ddH_2O.
Nach einer einleitenden Denaturierung der DNA für drei Minuten bei 90 °C folgen 35 Zyklen mit 30 Sekunden Denaturierung bei 96 °C, 30 Sekunden Annealing bei 55 °C und 45 Sekunden Elongation bei 72 °C. Der letzte Schritt umfasst eine zehn-minütige Elongation bei 72 °C, die gewährleistet, dass noch nicht vollständige DNA-Fragmente bis zum Ende der Matritze synthetisiert werden.
Die PCR-Produkte werden mittels Gelelektrophorese überprüft.

3.5 Reinigung von DNA aus Reaktionsansätzen

Bevor die amplifizierte DNA für weitere Analysen eingesetzt werden kann, muss sie von störenden Reaktionsüberresten wie der Polymerase, überschüssigen dNTPs, Primern, Salzen und Puffern gereinigt werden. Dazu wird das QIAquick PCR Purification Kit (Qiagen, Hilden) verwendet und nach Protokoll des Herstellers vorgegangen. Die gereinigten PCR-Fragmente werden bei -20 °C gelagert.

3.6 Sequenzierung

Zur Verifizierung von SNPs wird die klassische Sanger-Sequenzierung verwendet (Kapitel 3.6.1). Die Sequenzierung ganzer Genome wird dagegen unter Verwendung der beiden Sequenzier-Technologien der 2. Generation 454-Pyrosequenzierung und Solexa durchgeführt (Kapitel 3.6.2 und 3.6.3).
Tabelle 3.1 enthält eine Übersicht der Einrichtungen und Firmen, die die Sequenzierungen ausgeführt haben.

3.6.1 Sanger-Sequenzierung

Die enzymatische Sequenzierung nach Sanger und Coulson (Sanger, Nicklen & Coulson 1977) wird auch Kettenabbruch-Methode genannt. Neben dNTPs stehen der Polymerase auch Didesoxynukleotide (ddNTPs) zur Verfügung, deren Einbau aufgrund einer

Tabelle 3.1: Übersicht der verwendeten Sequenzier-Plattformen und -Einrichtungen.

Methode	Plattform	Einrichtung	Jahr
Sanger	3130xl Genetic Analyzer, Applied Biosystems, Darmstadt	RKI, Berlin	2008 - 2011
454-Pyrosequenzierung	GS 20, 454 Life Sciences, Branford, USA	454 Life Sciences, Branford, USA	2007
	GS FLX, Roche Diagnostic, Mannheim	RKI, Berlin	2009
	GS FLX, Roche Diagnostic, Mannheim	Georg-August-Universität, Göttingen	2010 - 2011
Solexa („*single-* & *paired-end*“)	Genome Analyzer, Illumina Inc., San Diego, USA	GATC, Konstanz	2006 - 2011
	HiSeq2000, Illumina Inc., San Diego, USA	GATC, Konstanz	2010 - 2011

fehlenden Hydroxylgruppe an der 3'-Position des Zuckers zum Abbruch der Strang-Synthese führt. Der 10 μL-Sequenzierungsansatz besteht aus folgenden Komponenten: 8,25 μL ddH_2O, 1 μL BigDye (Applied Biosystems, Darmstadt), 0,5 μL Primer und 0,25 μL PCR-Produkt.

3.6.2 454-Pyrosequenzierung

Die Pyrosequenzierung wurde 1996 von Pål Nyrén und Mostafa Ronaghi (Ronaghi, Uhlén & Nyrén 1989) entwickelt und durch 454 Life Sciences (heute: Roche Diagnostics) für Hochdurchsatz-Sequenzierungen kommerzialisiert. 454-Pyrosequenzierung ist die erste Sequenziertechnologie, die zur „*next generation*“ gezählt wird.
Der komplette Ablauf der 454-Pyrosequenzierung umfasst drei Hauptabschnitte, die im Folgenden kurz beschrieben werden.

Erstellung der DNA-Bibliothek. Die genomische DNA wird durch Zerstäubung (engl. „*nebulization*“) in 300 bis 800 bp lange, doppelsträngige Stücke fragmentiert. An die Enden werden zwei verschiedene Adapter ligiert, wobei ein Adapter biotinyliert vorliegt. Durch diesen können einzelsträngige Fragmente aussortiert werden, die für die folgenden Schritte benötigt werden.

Emulsions-PCR. Die ssDNA-Fragmente werden an Polystyrol-Kugeln (engl. „*capture beads*“) gebunden, die einzeln in einem Emulsionstropfen (Mikroreaktor) vorliegen, in dem die Amplifikation stattfindet. Die amplifizierten Fragmente werden an Adapter gebunden, die an der Oberfläche der Kugeln befestigt sind. Nach Abschluss der Emulsions-PCR ist die Oberfläche der Kugeln vollständig mit den gleichen Sequenzabschnitten überzogen.

Pyrosequenzierung. Nach der Emulsions-PCR werden die Mikroreaktoren aufgebrochen und jede Kugel in die einzelnen Vertiefungen auf einer PicoTiter-Platte geladen und die benötigten Sequenzier-Chemikalien (Enzyme, Nukleotide) hinzugefügt. Zur Pyrosequenzierung wird jeweils eines der vier Nukleotide über die Vertiefungen gespült. Sobald ein zum Ausgangsstrang passendes Nukleotid durch die DNA-Polymerase eingebaut wird, wird anorganisches Pyrophosphat (PP_i) abgespalten. Dieses wird von dem Enzym ATP-Sulfurylase in Gegenwart von Adenosin-5'-phosphosulfat (APS) zur Generierung von Adenosintriphosphat (ATP) benötigt. Das generierte ATP dient als Substrat für die durch Luciferase katalysierte Umwandlung von Luziferin zu Oxyluziferin bei der Energie in Form eines Lichtblitzes abgegeben wird. Die Intensität des Lichtblitzes ist dabei proportional zum ATP Verbrauch und gibt dadurch Auskunft wie oft ein Nukleotid eingebaut wurde. Der Blitz wird von einer CDD-Kamera detektiert, die alle Reaktionen verfolgt.

Im Gegensatz zu anderen NGS-Technologien hat 454 den großen Vorteil, relativ lange Reads zu generieren. Das neueste System GS FLX Titanum XL+ liefert Längen von bis zu 1000 bp (Durchschnitt: 700 bp; vgl. Durchschnitt für diese Arbeit: 400 bp) und kommt damit nah an Ergebnisse einer Sanger-Sequenzierung heran. Die größte Limitation der 454-Pyrosequenzierung liegt in der schlechten Auflösung von Homopolymeren. Da der Einbau von Nukleotiden nicht nach jedem Schritt abgebrochen und detektiert wird, steigt bei Homopolymeren nur die Intensität des Lichtblitzes. Aus dieser Intensität wird dann die Länge des Homopolymers abgeleitet (Stangier, Bauser & Regenbogen 2007, Shendure & Ji 2008).

3.6.3 Solexa/Illumina

Mitte der 1990er Jahre hatten S. Balasubramanian und D. Klenerman die Idee, kurze Reads parallel mithilfe einer Festphasen-Sequenzierung durch reversible Terminatoren zu sequenzieren („*sequencing-by-synthesis*“). In den folgenden Jahren gründeten sie die Firma Solexa (heute Illumina Inc.) und im Jahr 2006 wurde der erste Solexa-Sequenzierer auf den Markt gebracht (Illumina Inc. 2011).
Die drei Hauptschritte der Solexa-Sequenzierung werden im Folgenden kurz dargestellt.

Erstellung der DNA-Bibliothek. Die DNA wird fragmentiert und an den 3'-Enden ein A-Überhang angefügt. An diesen werden zwei verschiedene Adapter ligiert und die DNA anhand dieser selektiert.

Brücken Amplifikation. Die ssDNA wird über ihre Adapter an passende Oligonukleotide auf einer planaren, optischen Oberfläche (engl. „*flow cell*“) gebunden. Anschließend kommt es zur sogenannten „*bridge amplification*“, bei der ein antiparalleler Strang

synthetisiert wird, der wieder an ein Oligonukleotid bindet. Dadurch entstehen Cluster von bis zu 1.000 Fragmentkopien. Die brückenartig-gebogenen Amplifikate werden an einem Ende von der Oberfläche gelöst und es wird ein Sequenzierungsprimer an das freie Ende hybridisiert.

Sequenzierung. Für die Sequenzierung wird ein Nukleotidgemisch verwendet, in dem jedes der vier Nukleotide mit einem spezifischen Farbstoff markiert ist, der als 3'-Terminator dient. Sobald ein Nukleotid von der DNA-Polymerase eingebaut wird, wird der Replikationsvorgang gestoppt und die Fluoreszenz für jedes Cluster gemessen. Der Farbstoff wird entfernt und der Zyklus kann von vorne beginnen. Aus den aufgenommenen Bildern wird abschließend die Readsequenz generiert.

Neben der hier beschriebenen einfachen Solexa-Sequenzierung werden für diese Arbeit auch „*paired-end*"-Sequenzierungen durchgeführt. Dabei wird die ssDNA während der Brücken-Amplifikation erst vom einem Ende sequenziert, dann gedreht und vom anderen Ende erneut sequenziert. Dadurch werden Informationen über die Orientierung der Reads und dem Abstand zwischen den Reads bereitgestellt. Dies hat den großen Vorteil, dass Wiederholungsregionen, Rearrangements und Insertionen und Deletionen identifiziert werden können.

Obwohl für die Brücken-Amplifikation Fragmentlängen von 100 - 500 bp ideal sind, ist die Länge der sequenzierten Reads kleiner (diese Arbeit: 32 - 64 bp). Gründe dafür sind unter anderem die Abnahme der Fluoreszenz oder unvollständige Abtrennung der Farbstoffe. Im Gegensatz zur 454-Technologie werden Homopolymere korrekt aufgelöst (Shendure & Ji 2008). Durch ständige Verbesserung der Chemikalien sind mittlerweile Readlängen von bis zu 100 bp möglich.

4 Bioinformatische Methoden

4.1 Verarbeitung von Sequenzdaten

Die Rohdaten sowohl der 454-Pyrosequenzierung als auch der Solexa-Sequenzierung beruhen auf der Detektion von Licht- bzw. Fluoreszenzsignalen mittels einer Kamera und werden als einzelne Bilder abgespeichert. Dabei können pro Sequenzierung bis zu 28 Gb (454, Quelle: 454 Software Manual) bzw. 2 Tb (Solexa, persönliche Mitteilung Jochen Blom, CeBiTec Bielefeld) an Rohdaten generiert werden.
Da die Verarbeitung der Bild-Rohdaten sehr zeit- und rechenaufwändig ist, werden die Daten bereits in einem vorverarbeiteten Output, den sogenannten Reads, geliefert. Sequenzreads sind eine Abfolge von Basen mit - je nach verwendeter Sequenziertechnologie - unterschiedlicher Länge und stellen die Grundlage für die weitere Verarbeitung. Dabei kann zwischen zwei verschiedenen Verarbeitungstypen unterschieden werden:

- ***de novo* Assemblierung:** Zusammenfassung von überlappenden Reads in Contigs. Es entsteht eine neue, vorher unbekannte Sequenz.
- **Mapping:** Alignierung von Reads gegen eine Referenzsequenz. Sequenzbereiche, die nicht in der Referenz vorkommen, bleiben unerkannt.

Die nächsten Kapitel gehen näher auf die Weiterverwendung von Sequenzdaten und verwendete Methoden und Algorithmen ein.

4.1.1 Umwandeln von Sequenzdaten in das fastQ-Format

Neben der reinen Sequenzinformation eines Reads ist auch der zugehörige Qualitätswert eines sequenzierten Nukleotids für weitere Analysen wichtig. Der schon für die Sanger-Sequenzierung eingeführte PHRED-Qualitätswert (Q_{PHRED}; Ewing *et al.* 1998, Ewing & Green 1998) eines sequenzierten Nukleotids ist wie folgt definiert:

$$Q_{PHRED} = -10 * log_{10}(P_e) \quad (1)$$

mit P_e = Fehlerwahrscheinlichkeit einer identifizierten Base (engl. „*base calling error probability*“)

Während die Sequenzreads meist in einer *.fasta Datei gespeichert werden, werden die zugehörigen PHRED-Werte gesondert in einer *.qual Datei abgelegt.
Beide Dateien können in dem Sequenzformat fastQ zusammengefasst werden, welches sich in den letzten Jahren zu einem *de facto* Standard zur Datenspeicherung von verschiedenen „*next generation*“-Sequenziertechnologien entwickelt hat (Cock *et al.* 2009). Es setzt sich aus dem Sequenzread und dem zu jeder Base gehörigen Qualitätswert

zusammen, der anstelle eines numerischen Qualitätswertes als ASCII Zeichen codiert wird. Abbildung 4.1 zeigt ein Minimalbeispiel einer fastQ-Datei.

```
@NG-5015_09-03402_5_1_1_1162/1
ATAAGAATTTAACTAGTAAACTGAATGCGGT
+
?CABBABBBAAC@BBB3BA?@ABBB@B=BB<
```

Abbildung 4.1: Aufbau des fastQ-Formats.

Durch die schnelle Weiterentwicklung der verschiedenen Sequenziertechnologien kam es aber auch zur Entwicklung eigener, nicht mehr untereinander austausch- und vergleichbarer Qualitätswerte. Vor allem Illumina Inc. änderte mit jeder neuen Gerätegeneration die Definition der Qualitätswerte. Zur Zeit gibt es folgende fastQ-Varianten, die sich auch hinsichtlich der Umcodierung der Qualitätswerte in ASCII-Zeichen unterscheiden:

- Sanger, 454, Illumina 1.8
- Solexa/Illumina 1.0
- Illumina 1.3+
- Illumina 1.5

Um die verschieden Varianten, die auch für die Analysen dieser Arbeit vorliegen, vergleichbar zu machen, werden Perl-Skripte verwendet, die die Input-Dateien *.fasta und *.qual im Sanger-fastQ-Format zusammenfassen und die Qualitätswerte ($q) durch die folgenden Perlcodes in PHRED und ASCII ($Q) umwandeln:

PHRED-qual zu Sanger-fastQ:

```
$Q = chr($q + 33);
```

Solexa-qual zu Sanger-fastQ:

```
$Q = chr((10 * log(1 + 10 ** ($q/ 10.0))/log(10))+33);
```

Illumina1.3+-qual zu Sanger-fastQ:

```
$Q = chr($q + 64);
```

4.1.2 *De novo* Assemblierung von Sequenzdaten

Die Sequenzreads der Genomsequenzierungen werden mittels Assemblierungssoftware in Contigs zusammengefasst. Da die Länge der Reads stark von der verwendeten Sequenzierungsmethode abhängt, variieren auch die Anforderungen an die anzuwendenden Algorithmen.

Für die langen Reads des Pyrosequenzierers 454 GS FLX wird das Programm Newbler (Roche, Mannheim) benutzt. Die sehr viel kürzeren Solexa/Illumina Reads werden mit dem darauf spezialisierten Programm SOAPdenovo (Li *et al.* 2010) assembliert. Beide Programme werden mit den *default*-Einstellungen ausgeführt.

4.1.3 Mapping von Sequenzdaten

CC5. Da die Sequenzierungen der Isolate des klonalen Komplex CC5 hinsichtlich der verwendeten Methoden stark variieren (454 vs. Solexa/Illumina, „*single-end*“ vs. „*paired-end*“), wird das Programm SSAHA2 (Ning, Cox & Mullikin 2001) verwendet, um die selben Mapping-Bedingungen zu gewährleisten. SSAHA2 wird mit den *default*-Einstellungen ausgeführt und die Genomsequenz des Isolats N315 (Akzessionsnummer: BA000018.3) als Referenzsequenz verwendet.
Da das Outputformat *.sam sich nicht zur leichten Weiterverarbeitung eignet, wird mittels SAMtools (Li *et al.* 2009) aus der *.sam eine Datei im Pileup-Format generiert. Der Aufbau einer pileup-Datei wird in Kapitel 4.3.2 und Tabelle 4.1 näher beschrieben.

ST225. Für das Mapping der Isolate des Sequenztyps ST225 wird das Programm Maq (Li, Ruan & Durbin 2008) mit den *default*-Einstellungen genutzt.
Als Referenzsequenz dient das Genom 04-02981 (Akzessionsnummer: CP001844.2). Zur leichten Weiterverarbeitung wird das Outputformat Pileup gewählt.

4.2 Annotierung

Die Genome der Isolate werden mit Hilfe des Programms Kodon (Applied Maths, Sint-Martens-Latem, Belgien) auf der Grundlage der bereits in GenBank veröffentlichten *S. aureus*-Genome annotiert. Zur Bestimmung der „*open reading frames*“ (ORFs) wird der genetische Code für Bakterien (Translationstabelle 11) verwendet.
ORFs ab einer Größe von 99 bp (individuell festlegbar) werden angezeigt und annotiert. In den Referenzgenomen nicht vorkommende Gene werden mit der BLAST-Methode blastn mit den in GenBank enthaltenen Genen abgeglichen und bei einer Homologie von mindestens 50 % annotiert. ORFs, die keine Homologien mit bisher bekannten Genen aufweisen und kleiner als 300 bp sind, werden gelöscht.

4.3 Erstellen von Alignments

Ein Alignment stellt eine wichtige Grundlage für viele weitere Analysen wie z.B. Genomvergleiche oder phylogenetische Analysen dar. Je nach Weiterverwendung können sich die verwendeten Alignment-Methoden stark unterscheiden.

4.3.1 Alignments für Genomvergleiche

Für die Durchführung von Genom-basierten Alignments werden die Programme Kodon (Applied Maths, Sint-Martens-Latem, Belgien) und Mauve (Darling, Mau & Perna 2010) verwendet.
Kodon hat gegenüber anderen Programmen, die Genome alignieren können, den Vorteil einer sehr guten grafischen Darstellung. Somit sind Vergleiche der Genome, die z.B. die genetische Ausstattung der untersuchten Organismen betreffen, sehr leicht durchführbar (Kapitel 5.2.5).
Mauve (Darling, Mau & Perna 2010) verwendet im Gegensatz zu Kodon einen nicht Referenzsequenz-basierten Algorithmus. Dadurch ist es besonders für die Konstruktion von Alignments von Genomen geeignet, die starken Umstrukturierungen (z.B. Rearrangements, Inversionen) unterworfen sind. In dieser Arbeit wird Mauve für die Alignierung von Prophagen-Genomen verwendet (Kapitel 4.11.1).

4.3.2 Alignments für phylogenetische Analysen

Die Rekonstruktion der Phylogenie des klonalen Komplexes CC5 und des Sequenztyps ST225 soll unter Verwendung ganzer Genomsequenzen durchgeführt werden. Um mögliche Sequenzierfehler auszugleichen, die später zu falschen evolutionären Zusammenhängen führen könnten, sollen auch die Qualitätswerte der eingebauten Basen berücksichtigt werden. Das genaue Vorgehen zum Aufbau eines Qualitäts-geprüften Kerngenom-Alignments wird im Folgenden näher beschrieben. Abbildung 4.2 zeigt die notwendigen Schritte in der Übersicht.

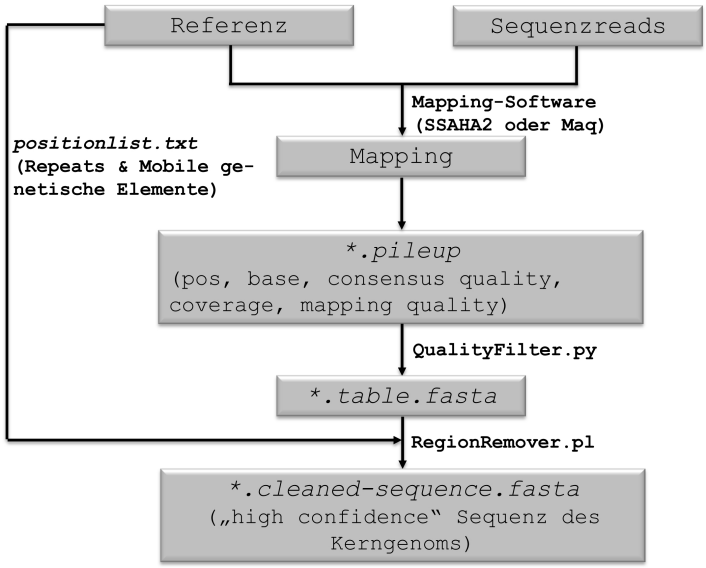

Abbildung 4.2: Ablaufdiagramm zur Erstellung von Qualitäts-geprüften Kerngenom-Sequenzen.

Identifizierung von Wiederholungssequenzen und mobilen genetischen Elementen. Kommen Sequenzabschnitte in einem Genom mehrfach vor, kann es zu einem falschen Mapping der Reads kommen, da diese zufällig an einen Abschnitt aligniert werden. In einem solchen Fall würde man mit einer falschen Sequenz arbeiten. Aus diesem Grund werden Regionen mit Wiederholungssequenzen bei phylogenetischen Analysen nicht berücksichtigt.
Zur Identifizierung von Wiederholungssequenzen im Referenzgenom wird das Programm Kodon verwendet. Das Ergebnis ist eine Liste mit Start- und Endposition der Wiederholungssequenzen.
Neben Repeats sollten auch mobile genetische Elemente wie Prophagen oder Pathogenitätsinseln nicht berücksichtigt werden, da diese oft horizontal übertragen werden. Dadurch würde der Einschluss von Sequenzunterschieden in diesen Elementen zur Rekonstruktion einer falschen Phylogenie führen. Die verwendeten Referenzgenome N315 und 04-02981 sind bereits beschrieben und annotiert (Kuroda *et al.* 2001, Nübel *et al.* 2010) und somit können die Start- und Endpositionen der mobilen genetischen Elemente aus diesen übernommen werden.
Die Positionen der Wiederholungssequenzen und der mobilen genetischen Elemente werden untereinander in einer *.txt Datei (positionlist.txt) gespeichert und in einem späteren Schritt weiterverwendet.

Mapping. Auf die verwendeten Programme und Parameter zum Mappen von Sequenzreads auf eine Referenz wurde bereits in Kapitel 4.1.3 näher eingegangen. Als Output erhält man u.a. eine Tabelle, die neben der Position im Referenzgenom auch die sequenzierte Base und verschiedene Qualitätswerte des Mappings enthält. Tabelle 4.1 zeigt den genauen Aufbau einer solchen Datei im Pileup-Format. Obwohl sich diese zwischen SSAHA2 und Maq leicht unterscheiden (z.B. gibt Maq keine SNP Qualität an), wird hier zum leichteren Verständnis nur auf die SSAHA2 *.pileup Datei eingegangen. Der weitere Ablauf ist aber nahezu identisch.

Tabelle 4.1: Aufbau einer SSAHA2 *.pileup Datei.

Referenz	Position in der Ref.	Referenz Base	Konsensus Base	Konsensus Qualität	SNP Qualität	Mapping Qualität	Coverage
N315	1	C	C	147	0	35	56
N315	2	A	A	150	0	35	2
N315	3	T	A	16	50	35	16
N315	4	*†				35	56
N315	5	G	G	10	0	1	137
N315	6	G	G	130	0	29	125

† an dieser Position befindet sich ein Indel, was durch das Zeichen * angezeigt wird

Filtern nach Qualitätswerten. Da es bei der Sequenzierung zu Fehlern kommen kann, sollen verschiedene Qualitätswerte genutzt werden, um Qualitäts-geprüfte Sequenzen zu erhalten. Für diesen Filterschritt wurde das Python-Skript „QualityFilter.py“ geschrieben und verwendet. Dieses liest die *.pileup Datei zeilenweise ein und schreibt alle Basen, die unter einen frei zu wählenden Qualitäts-Grenzwert fallen, in ein „N“ um. Das „N“ steht dabei nach der IUPAC Nomenklatur für Nukleinsäuren für eine unspezifische Nukleobase. Abschließend werden zwei Dateien erstellt: die erste enstpricht der *.pileup-Datei mit den veränderten Basen (*.table) und die zweite enthält die Sequenz im Fasta-Format (*.table.fasta).
Tabelle 4.2 gibt eine Übersicht über die verwendeten Grenzwerte der verschiedenen Qualitätswerte. Der Output, der unter diesen Bedingungen für eine Input-Datei (Tabelle 4.1) generiert werden würde, ist in Abbildung 4.3 gezeigt.

Tabelle 4.2: Grenzwerte der verschiedenen Qualitäten.

Konsensus Qualität	SNP Qualität	Coverage	Mapping Qualität
SSAHA2 (CC5)			
30	20	3	30
Maq (ST225)			
60	-	3	40

A

```
N315   1   C   C   147    0   35    56
N315   2   A   N   150    0   35     2
N315   3   T   A    16   50   35    16
N315   5   G   N    10    0    1   137
N315   6   G   G   130    0   29   125
```

B

```
>Bsp
CNA-NG
```

Abbildung 4.3: Output nach Qualitätsfilterung. A: Pileup-Format (*.table). B: Fasta-Format (*.table.fasta)

Entfernen von Wiederholungssequenzen und mobilen genetischen Elementen. Um Wiederholungssequenzen und mobile genetische Elemente aus der Qualitätsgefilterten Sequenz zu entfernen wurde das Perl-Script „RegionRemover.pl“ geschrieben und verwendet. Dieses liest die Sequenz im Fasta-Format (*.table.fasta) und die im ersten Schritt erstellte Textdatei (positionlist.txt) ein, die die Start- und Endpositionen der Wiederholungssequenzen und mobilen genetischen Elementen enthält. Diese Regionen werden aus der Sequenz entfernt und es wird eine neue *.fasta Datei (*.cleaned-sequence.fasta) erstellt.

Zusammenstellen der Sequenzen. Die nun vorliegenden Sequenzen enthalten nach Qualitätsfilterung und Entfernung von Wiederholungssequenzen und mobilen genetischen Sequenzen nur noch Qualitäts-geprüfte Nukleotide. Die gewünschten *.fasta-Dateien werden in MEGA (Tamura *et al.* 2011) eingeladen und können dort leicht in andere Formate konvertiert oder weiterverarbeitet werden (z.B. Ausschluss von nicht aufgelösten Nukleotiden und Lücken, Extrahieren von variablen Alignment-Positionen). Nun liegt ein fertiges Alignment der gewünschten Sequenzen vor, mit dem die phylogenetischen Analysen durchgeführt werden.

4.4 Phylogenetische Analysen

Da phylogenetische Analysen mit ganzen Genomsequenzen sehr rechenintensiv sind, werden mit MEGA5 (Tamura *et al.* 2011) die variablen Merkmale aus den zu verwendenden Alignments extrahiert und mit diesen gearbeitet.

4.4.1 „*Maximum Likelihood*“ Analysen

Bei der „*Maximum Likelihood*“ (ML)-Methode wird der Stammbaum gesucht, der den höchsten Likelihood-Wert besitzt. Der Likelihood-Wert L ist dabei die Wahrscheinlichkeit, dass die ermittelten Daten D (Merkmale im Alignment) aufgrund der Hypothese H (Stammbaum und Evolutionsmodell) entstanden sein können (Schmitt 2006).

Bestimmung des optimalen Substitutionsmodells. Soll die phylogenetische Analyse auf der „*Maximum Likelihood*“-Methode beruhen, ist es dringend notwendig, vorher das optimale Substitutionsmodel zu bestimmen (Posada & Crandall 1998). Substitutionsmodelle machen Annahmen über den Verlauf von Sequenzevolution (Knoop & Müller 2006) und werden in Treefinder (Jobb, von Haeseler & Strimmer 2004) ermittelt.
Wie gut das Modell zu den vorgegebenen Daten passt, kann durch die Likelihood des Modells bestimmt werden. Je größer die Anzahl der zum Modell zugefügten Parameter ist, desto größer wird die Varianz der erhaltenen Werte. Der zu erwartende Fehler kann durch das Akaike Informationskriterium (AIC, Akaike 1974) abgeschätzt werden. Das beste Modell wird durch das geringste korrigierte AIC (AICc) angegeben, das nicht wie das AIC nur die Anzahl der Modell-Parameter, sondern auch die Anzahl der Merkmale berücksichtigt.

„*Maximum Likelihood*“ Analyse. Die „*Maximum Likelihood*“-Analyse wird in Treefinder (Jobb, von Haeseler & Strimmer 2004) durchgeführt. Dazu wird ein Alignment eingeladen und das vorher bestimmte Substitutionsmodell ausgewählt. Zusätzlich wird die Unterstützung der Knoten berechnet.

Statt Bootstrap (BS)-Werten werden in Treefinder so genannte LR-ELWs berechnet. Dabei handelt es sich um approximierte Bootstrap-Werte, die meist etwas höher ausfallen als gewöhnliche Bootstrap-Werte. Es wird die Standard-Einstellung mit 1.000 Wiederholungen beibehalten. Sobald die Analyse abgeschlossen ist, wird der Baum im Newick-Format abgespeichert.

4.4.2 Bayes'sche Analysen

Bayes'sche Analysen beruhen auf der so genannten „*posterior probability*“, die basierend auf der Wahrscheinlichkeit einer anfänglichen Hypothese („*prior probability*“) und neuen Erkenntnissen nach einem Experiment ermittelt wird.
Der Unterschied zur „*Likelihood*“-Analyse ist, dass nicht die Wahrscheinlichkeit der Daten berechnet wird, sondern die Wahrscheinlichkeit der Hypothese aufgrund der Daten. Die Bayes'sche Analyse sucht also die Bäume mit den höchsten „*posterior probability*“-Werten.
Für die Rekonstruktion des Stammbaums wird MrBayes (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003) verwendet. Als Substitutionsmodell wird das „*general time reversible*“-Modell mit Γ-Verteilung und einem Anteil invarianter Positionen (GTR+G+I) mit dem Befehl `lset nst= 6 rates= invgamma;` festgelegt. Die „*prior probabilities*“ werden mit `prset revmatpr= dirichlet(1,1,1,1,1,1) statefreqpr= dirichlet(1,1,1,1) shapepr= uniform(0.1,50) pinvarpr= uniform(0,1);` bestimmt.
Für die Berechnung wird der „*Metropolis-coupled Markov Chain Monte Carlo*“ (MCMCMC)-Algorithmus angewendet. Es laufen zwei parallele Berechnungen mit je vier Ketten. Insgesamt werden 2.250.000 Generationen durchlaufen und die Bäume jeder 250sten Generation mit Astlängen in einer Datei gespeichert. Der dazugehörige Befehl lautet: `mcmcp ngen= 225000000 nchains= 4 samplefreq= 100 printfreq= 250 savebrlens=yes Diagnfreq= 2500;`. Die Analyse startet mit `mcmc`.
Nachdem die angegebene Anzahl der Generationen abgearbeitet ist, kann aus allen gespeicherten Stammbäumen ein 50 %-„*Majority Rule*“-Stammbaum gebildet werden. Da die Ketten eine gewisse Zeit brauchen, um sich einzupendeln, werden nur die Bäume nach diesem „*burn-in*“ verwendet. Um den Zeitpunkt des „*burn-in*“ zu ermitteln, wird ein Diagramm erstellt, in dem die Standardabweichung der „*Likelihood*“-Werte gegen die Generationen aufgetragen werden (Abbildung 4.4). Sobald die „*Likelihoods*“ nicht mehr sinken und um einen Wert schwanken, haben sie sich stabilisiert. Mit dem Befehl `sumt burnin= x` werden alle Bäume vor dem „*burn-in*“ ignoriert und aus den übrigen der Konsensusbaum gebildet und gespeichert. Tritt der Fall ein, dass sich die Ketten nach der angegebenen Anzahl an Generationen noch nicht im Equilibrium befinden, wird die Analyse fortgeführt.

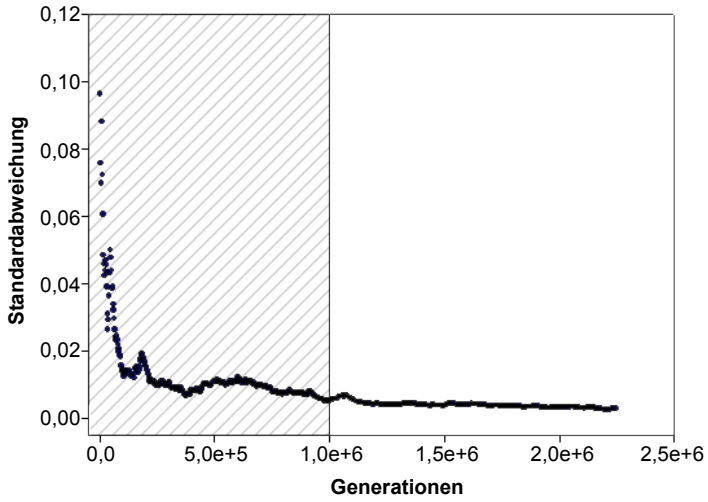

Abbildung 4.4: Gleichgewichtsbestimmung der Bayes'schen Analyse. Gezeigt ist die Standardabweichung der „*Likelihood*"-Wert der gespeicherten Topologie jeder 250sten Generation gegen die Generationen. Der schraffierte Bereich zeigt die Anzahl der Generationen vor dem „*burn-in*" und wird verworfen.

4.5 Berechnung von Substitutionsraten und Zeiten

Zur Berechnung von Substitutionsraten und Datierung von Entstehungszeiträumen des Sequenztyps ST225 wird BEAST (Drummond & Rambaut 2007) verwendet. BEAST beruht auf Bayes'scher Statistik und verwendet den „*Markov Chain Monte Carlo*" (MCMC)-Algorithmus. Da für die Datierung von bakteriellen Divergenzzeitpunkten keine fossilen Belege vorhanden sind, können in BEAST zur Kalibrierung der molekularen Uhr die Zeitpunkte der Probenentnahme berücksichtigt werden.
Als Input für BEAST muss zuerst mit der begleitenden Software BEAUti (Drummond, Rambaut & Suchard 2010) eine XML-Datei erzeugt werden, die neben dem Alignment auch die Einstellungen zum Substitutionsmodell, dem Modell der molekularen Uhr und den „*priors*" auch die Daten der Probennahme enthält. Tabelle 4.3 enthält eine Übersicht über die verwendeten Einstellungen. Die XML-Datei wird in BEAST eingeladen und ausgeführt. Um sicher zu stellen, dass die Analyse nicht in einem lokalen Optimum feststeckt, wird jede BEAST-Analyse zweimal durchgeführt und die Konvergenz der Ergebnisse mit Tracer (Rambaut & Drummond 2007) kontrolliert. Sind die nach dem MCMC-Lauf berechneten „*effective sample size*" (ESS) Parameter größer als 300, ist die Kette lange genug gelaufen und es wurden genügend Generationen berechnet; die Ergebnisse sind zuverlässig.

Tabelle 4.3: Einstellungen für Bayes'sche Analyse mit BEAST.

Einstellung	Parameter
Site Models	
Substitution Models	HKY
Base frequencies	Estimated
Site heterogeneity Model	None
Partition into codon positions	Off
Clock Models	
Model	strict
Rate	1.0
Trees	
Tree Prior	Constant size
Priors	*default*
Operators	*default*
MCMC	
Length of Chain	variieren
Echo state to screen every	variieren
Log parameters every	variieren

4.6 Berechnung von Homoplasien

Bei Homoplasien handelt es sich um identische Merkmalszustände, die sich nicht auf einen gemeinsamen Ursprung zurückführen lassen, sondern durch konvergente Evolution oder Rekombination entstanden sind.
Der Grad an Homoplasien wird mit dem Programm PAUP* (Wilgenbusch & Swofford 2003) ermittelt. Ein Homoplasie-Index (HI) von 1 bedeutet, dass alle Merkmale homoplastisch sind, wogegen ein HI von 0 auf homologe Evolution hindeutet. Die Positionen von homoplastischen SNPs im Datensatz werden mit dem Programm DnaSP (Librado & Rozas 2009) bestimmt und mittels PCR überprüft. Die verwendeten Primer sind im Anhang in Tabelle A.1 aufgeführt.

4.7 Berechnung von dN/dS

Nukleotidsubstitutionen in einem Gen können zu einer Änderung der kodierten Aminosäure führen; in diesem Fall spricht man von einer nicht-synonymen Mutation. Aufgrund der Degeneration des genetischen Codes kann es aber auch vorkommen, dass die Aminosäure erhalten bleibt (synonyme Mutation). Diese Tatsache erlaubt es unter gewissen Umständen, eine Aussage über das Maß von Selektionsdruck auf ein Gen zu treffen. Dazu wird das normalisierte Verhältnis von nicht-synonymen (dN) zu synonymen (dS) Nukleotidsubstitutionen mittels folgender Formel nach Nei & Gojobori (1986) berechnet:

$$dN/dS = \frac{(n/N)}{(s/S)} \tag{2}$$

mit n = Summe der nicht-synonymen SNPs
N = nicht-synonyme Seiten
s = Summe der synonymen SNPs
S = synonyme Seiten

Werden in einem Gen mehr nicht-synonyme Mutationen gefunden, als sich durch zufällige Mutationen erklären lassen, ist $\mathrm{d}N/\mathrm{d}S > 1$ und man spricht von positiver oder diversifizierender Selektion. Dabei verändert sich das Protein so, dass es besser an äußere Gegebenheiten angepasst ist. Ist $\mathrm{d}N/\mathrm{d}S \ll 1$ wird von stabilisierender oder reinigender Selektion gesprochen und schädliche Mutationen werden aus dem Erbgut entfernt. Bei einer zufälligen Entstehung von synonymen und nicht-synonymen Mutationen ist $\mathrm{d}N/\mathrm{d}S = 1$ und es handelt sich um neutrale Evolution.
Für die Berechnung von $\mathrm{d}N/\mathrm{d}S$ wird die Software DnaSP (Librado & Rozas 2009) verwendet.

4.8 Berechnung der Häufigkeit AT-anreichernder Mutationen

Die Berechnungen zur Häufigkeit AT-anreichernder Mutationen werden Balbi, Rocha & Feil 2008 folgend durchgeführt.
Zuerst wird anhand der Phylogenie des zu untersuchenden Datensatzes die Richtung der Mutationen bestimmt (z.B. X → Y oder Y → X). Um die Häufigkeit der zwölf möglichen Mutationsmöglichkeiten zu berechnen, wird die Anzahl jedes Typs gezählt und durch die Zahl der betrachteten Originalbase geteilt (Gojobori, Ishii & Nei 1982).
Beispiel. Die Häufigkeit einer C → T Transition in Genom 1 wird in zwei Schritten berechnet: Zuerst wird die Anzahl der C → T Mutationen im Genom 1 gezählt. Diese Zahl wird dann durch die Gesamtzahl an „C“s im Genom 1 geteilt.
Die Häufigkeit aller zwölf Mutationsmöglichkeiten wird auf 1 normalisiert, um jede Möglichkeit als Anteil aller Mutationen in einem Genom betrachten zu können und die ungleiche Verteilung von GC bzw. AT Basen im Genom zu berücksichtigen. Vom Polymorphismus-Profil ausgehend wird eine Rate berechnet, die der Summe der normalisierten Häufigkeiten der AT → GC Mutationen geteilt durch die normalisierten Häufigkeiten der GC → AT Mutationen (+GC/+AT) entspricht.
Diese Rate wird sowohl für jedes einzelne Genom als auch für die ursprünglichen[1] Äste berechnet. Um das Alter der Äste zu berechnen, wird die Anzahl synonymer

[1] als „ursprünglich“ werden im phylogenetischen Kontext die innenliegenden und damit älteren Verzweigungen eines Stammbaums bezeichnet. Im Englischen wird das Wort „*ancestral*“ verwendet. Das Gegenteil sind die Spitzen im Stammbaum, die als „terminal“ bezeichnet werden.

Mutationen berücksichtigt. Für die terminalen Äste werden einfach die synonymen SNPs gezählt; für die ursprünglichen Äste wird die Zahl der synonymen SNPs auf dem betrachteten Ast durch zwei geteilt und mit dem Mittelwert der synonymen SNPs vom letzten Knoten bis zu den terminalen Ästen addiert.

4.9 Zuordnen von Genkategorien

Um zu überprüfen, ob bestimmte Genkategorien häufiger von SNPs betroffen sind als andere, werden die Gene der Genome des klonalen Komplexes CC5 in funktionelle Klassen eingeteilt. Als Grundlage dafür wird die Kategorisierung der N315-Gene verwendet, die von der „Comprehensive Microbial Resource“ (CMR) des J. Craig Venter-Instituts bereitgestellt wird.

4.10 Prophagen im *S. aureus* Genom

4.10.1 Prophagen in 454 Sequenzen

Die Identifizierung von Prophagen in den mit 454 sequenzierten *S. aureus* Genomen erfolgt anhand der Integrase-Sequenz. Dazu wird nach den Primersequenzen gesucht, die von Goerke *et al.* (2009) zur Klassifizierung von *S. aureus* Prophagen entwickelt wurden (Kapitel 4.10.3). Die großen Contigs der 454-Sequenzierung haben den Vorteil, dass komplette Prophagen meistens in einzelnen Contigs enthalten sind. In den wenigen Fällen in denen dies nicht der Fall ist, werden die Contigs mittels Abacas (Assefa *et al.* 2009) anhand verschiedener Referenzen sortiert und so die Prophagen zusammengesetzt.

4.10.2 Prophagen in Solexa Sequenzen

Aufgrund der kurzen Reads, die bei einer Solexa-Sequenzierung entstehen, ist eine Identifizierung von Prophagen in solchen Sequenzen schwerer durchzuführen. Mit dem normalen Vorgehen, die Reads gegen eine Referenz zu mappen, kann hier nicht gearbeitet werden. Regionen (in diesem Fall verschiedene Prophagen), die in der Referenz nicht vorkommen, würden nicht erkannt werden.
Abbildung 4.5 zeigt das im Weiteren näher beschriebene Vorgehen in der Übersicht.

Mapping. Die Reads einer Sequenzierung werden mit Maq (Li, Ruan & Durbin 2008) gegen das Genom des Isolats N315 aligniert, aus dem die Sequenz des Prophagen ϕN315 vorher entfernt wurde. Dazu wird der Befehl `maq match -u outnir.txt out.map reference.bfa reads.bfq` verwendet; die Option `-u` gibt die Reads aus, die nicht in der Referenz vorkommen.

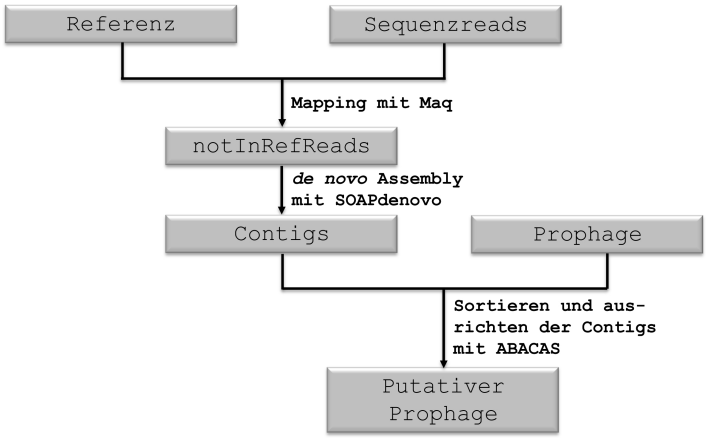

Abbildung 4.5: Ablaufdiagramm zur *de novo* Assemblierung von Prophagen.

***de novo* Assemblierung.** Für die *de novo* Assemblierung der nicht in der Referenz vorkommenden Reads wird SOAPdenovo-31mer (Li *et al.* 2010) verwendet, welches speziell für die kurzen Illumina GA Reads entwickelt wurde. Der Algorithmus basiert auf der „de Bruijn"-Methode und ist durch die Verwendung von K-mer Graphen besonders schnell. Ein weiterer Vorteil ist, dass für kleine Datensätze (wie Bakterien-Genome) ein primärer Filterschritt (z.B. zum Aussortieren von Reads mit schlechten Qualitätswerten) nicht nötig ist, da fehlerhafte Graphenverbindungen während der Assemblierung noch korrigiert werden (Li *et al.* 2010). SOAPdenovo-31mer wird mit den Standard-Einstellungen (K-mer 23) ausgeführt.

Sortieren der Contigs. Um aus den Contigs Prophagen-Sequenzen zu generieren, müssen sie noch anhand einer Referenz in die richtige Reihenfolge und Orientierung gebracht werden. Als Referenz dienen 79 verschiedene *S. aureus* Prophagen-Genome, die entweder aus GenBank entnommen oder aus den vorhandenen Genomen extrahiert wurden. Die Contigs werden nun abschließend mit Abacas (Assefa *et al.* 2009) einzeln gegen jeden der 79 Prophagen sortiert. Die Sequenz mit den wenigstens Contigs, einer passenden Prophagen-Genomgröße von ca. 44 kb und einem intakten und vorhandenen Integrase-Gen wird für die weiteren Analysen verwendet.

4.10.3 Klassifizierung von Prophagen

Zur Klassifizierung der Phagen werden drei verschiedene Ansätze verwendet.

Integrase-Sequenzen. Eine erste, grobe Einteilung der Prophagen basiert auf der Sequenz ihres Integrase-Gens und beruht auf einer Idee von Goerke *et al.* (2009). Die Integrase eignet sich laut Goerke *et al.* aus verschiedenen Gründen besonders gut: 1. die Nukleotid-Sequenz des Gens ist konserviert, 2. große diskriminatorische Stärke, die die Diversität der *S. aureus*-Prophagen ebenso gut aufzeigt wie ihre Verwandtschaft

und 3. der Integrase-Typ ist verknüpft mit im Genom enthaltenen Virulenzgenen. Zur Identifizierung des Integrase-Typs werden die in Goerke *et al.* (2009) beschriebenen Primer verwendet und *in silico* in den Genomsequenzen nach den Primern gesucht.

Nukleotid-Sequenzen. Die zweite Klassifizierung beruht auf dem gesamten Prophagen-Genom auf Nukleotidebene. Eine Clusterung erfolgt mittels UPGMA (Sokal & Michener 1958, Murtagh 1984) in Kodon (Applied Maths, Sint-Martens-Latem, Belgien).

Protein-Sequenzen. Die dritte und aufwendigste Klassifizierung basiert auf den Proteinen der Prophagen-Genome und folgt Leplae *et al.* (2004). Zuerst werden paarweise Ähnlichkeiten zwischen den Proteinsequenzen mittels BlastP (Altschul *et al.* 1990) und einem E-Wert Grenzwert von 10^{-20} erfasst. Die Einteilung in Protein-Familien erfolgt dann mittels des Markov Cluster Algorithmus (MCL) (Enrigth, Van Dongen & Ouzounis 2002), der in BioLayout *Express*^{3D} (Theocharidis *et al.* 2009) implementiert ist, unter Berücksichtigung eines Inflation-Werts von I=1,5.
Die Zusammensetzung der Protein-Familien innerhalb der Prophagen wird in einer Matrix zusammengefasst, in der die Spalten den Protein-Familien und die Zeilen den Prophagen-Genomen entsprechen. Das Vorkommen einer Protein-Familie wird dabei mit einer 1 und das Fehlen mit einer 0 codiert. Diese binäre Matrix wird als Input für PAUP* (Wilgenbusch & Swofford 2003) zur Berechnung einer Distanzmatrix verwendet, mit der im nächsten Schritt ein gewurzelter UPGMA-Baum (Sokal & Michener 1958, Murtagh 1984) rekonstruiert wird.

4.11 Detektion von Rekombinationsereignissen in Prophagen

Wie bereits in Kapitel 4.10.3 *Integrase-Sequenzen* beschrieben, können die Prophagen in verschiedene Integrase-Gruppen eingeteilt werden. Die folgenden Analysen werden für jede Integrasegruppe gesondert durchgeführt.

4.11.1 Mauve Alignment

Als Input für die Programme zur Detektion von Rekombinationsereignissen wird ein Alignment benötigt, welches mit Mauve (Darling, Mau & Perna 2010) erstellt und manuell überprüft wird. Da die Identifizierung von Rekombination auf Unterschieden zwischen den Sequenzen beruht, ist ein zu hohes Maß an Sequenzähnlichkeit nicht erwünscht. Trotzdem sollen nicht zu viele Lücken das Alignment künstlich ausdehnen. Aus diesem Grund werden die Straf-Parameter für das Einfügen einer Lücke im Alignment („*gap opening penalty*“) und das Ausweiten einer Lücke („*gap extension penalty*“) extrem hoch gesetzt (4000 & 3999).

4.11.2 ClonalFrame

Zur Identifizierung von Rekombinationsereignissen wird ClonalFrame genutzt (Didelot & Falush 2007). ClonalFrame wurde speziell zur Rekonstruktion von Verwandtschaftsbeziehungen rekombinanter Organismen entwickelt, da für diese konventionelle phylogenetischen Methoden nicht angewendet werden können. Für diese Arbeit sind vor allem die Informationen über die Wahrscheinlichkeit von Rekombinationsereignissen im Allgemeinen und an bestimmten Genomposition von Interesse.

ClonalFrame Output. Die Output-Datei von ClonalFrame umfasst 13 Teile von denen für diese Arbeit aber nur die Teile `#consevents` und `#poly` von Bedeutung sind.
`#consevents` enthält Informationen über gefundene Rekombinationsereignisse. Der Aufbau dieses Abschnitts entspricht einer N x (2M) Matrix mit N = Anzahl der Knoten im Baum und M = Anzahl der Referenzmerkmale (siehe `#poly`) mit 1. Wahrscheinlichkeit, dass Unterschiede an dieser Position durch Rekombination entstanden sind und 2. Wahrscheinlichkeit, dass Unterschiede an dieser Position durch Substitution entstanden sind (Rekombination oder Mutation). `#poly` enthält die sogenannten Referenzmerkmale.
Eine ausführlichere Beschreibung des Outputs ist im Manual von ClonalFrame zu finden (`http://www.xavierdidelot.xtreemhost.com/clonalframe.htm`).

Vorbereitung des Outputs für weitere Analysen. Da nur die Wahrscheinlichkeit von Rekombination im Genom von Interesse ist, wird die Spalte, die die Substitutionswahrscheinlichkeit enthält, entfernt und die Matrix transponiert. Die Zeilen der Matrix entsprechen nun den Zeilen der Referenzmerkmale. Die Referenzmerkmale werden zusammen mit den Rekombinationswahrscheinlichkeiten eines Knotens in einer *.txt-Datei abgespeichert. Bei N = 5 liegen am Ende also fünf Textdateien vor.

Bestimmung von rekombinierten Segmenten. Um den Start und das Ende eines rekombinierten Segments zu bestimmen, wird ein selbstgeschriebenes Perl-Programm verwendet („ClonalFrameParser“), welches die *.txt-Dateien einliest. Sobald die Wahrscheinlichkeit von Rekombination $\geq$ 0,95 ist, beginnt ein Segment und endet sobald die Wahrscheinlichkeit unter 0,95 sinkt. Der Grenzwert wurde Didelot *et al.* (2009) folgend definiert. Segmente, die die oben beschriebene Bedingung zwar erfüllen, aber nur eine Länge von 1 bp haben, werden nicht berücksichtigt, da es sich bei diesen Ereignissen eher um SNPs als um rekombinierte DNA handelt. Der Output enthält jeweils den Start und das Ende eines rekombinierten Fragmentes, wobei die Positionen den Positionen im Alignment entsprechen.

Abbildung der Rekombinationsereignisse auf die CC5 Phylogenie. Die Rekombinationsereignisse der Prophagen werden nach dem Parsimonie-Prinzip auf die Phylogenie des bakteriellen Wirts (hier: Isolate des klonalen Komplexes CC5) abgebildet. Verwandtschaftsbeziehungen zwischen den bakteriellen Genomen, die identische Prophagen-Fragmente enthalten, werden dabei berücksichtigt. So kann eine Aussage über den Zeitpunkt gemacht werden, wann in der Evolution Rekombination im Prophagen stattgefunden hat.

Korrelation zwischen Bakterien- und Prophagendiversität. Ein Mechanismus, der die große Diversität zwischen Prophagengenomen erklären kann, ist unbekannt. Um einen Hinweis auf diesen zu bekommen, wird die paarweise Distanz zweier Bakterien mit der Anzahl der Rekombinationsereignissen, die zwischen ihnen liegen, korreliert. Die Distanz zwischen den bakteriellen Genomen wird mit MEGA4 (Tamura *et al.* 2007) berechnet. Zur Bestimmung der Signifikanz einer Korrelation wird der Mantel-Test (Mantel 1967) verwendet, der in der Software zt (Bonnet & Van de Peer 2002) implementiert ist.

5 Ergebnisse

5.1 Das Genom 04-02981

Das Genom 04-02981 des Sequenztyps ST225 wurde im Rahmen dieser Arbeit mit den beiden NGS-Technologien 454 und Solexa sequenziert und mittels long-PCR und Sanger-Sequenzierung geschlossen. Es ist unter der Akzessionnummer CP001844 in GenBank hinterlegt.
Abbildung 5.1 zeigt eine zirkuläre Darstellung des Chromosoms.

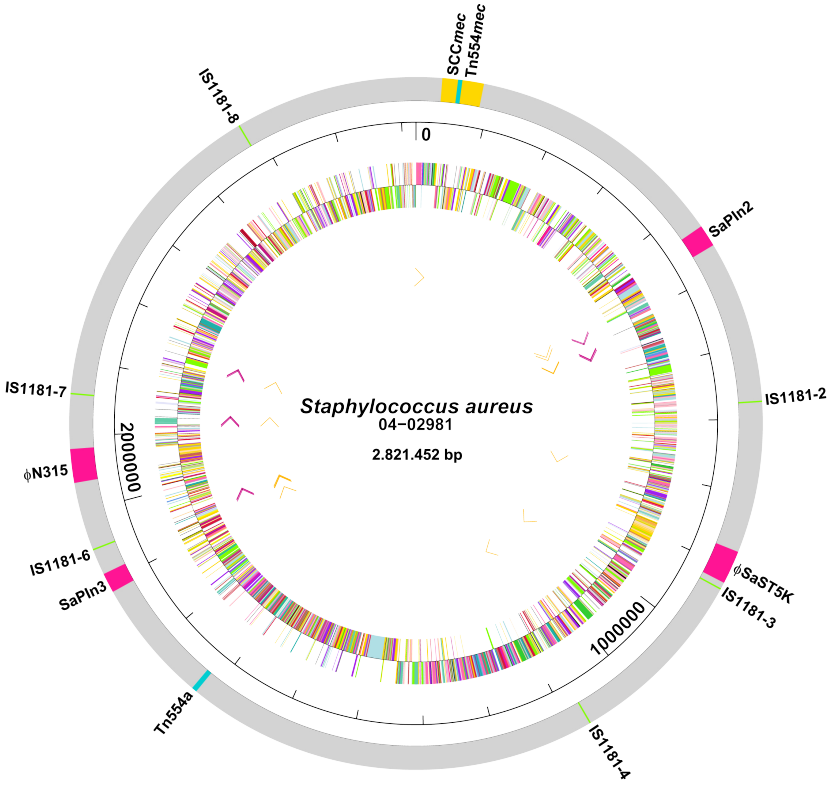

Abbildung 5.1: Zirkuläre Darstellung des Genoms 04-02981. Die Ringe geben von außen nach innen folgende Merkmale wieder: Mobile genetische Elemente (gelb: SCC*mec*, pink: Pathogenitätsinseln und Prophagen, grün: Insertionselemente, blau: Transposons); Nucleotidposition in bp; Gene auf dem + und dem - Strang; rRNAs; tRNAs

5.1.1 Sequenzierung

454. Die Assemblierung von 391.703 454-Reads mit Newbler ergibt 42 Contigs mit einer Größe $>$ 500 bp. Die Anzahl der enthaltenen Basen beträgt 2.774.621, was einer Abdeckung von 98 % des Genoms entspricht. Die durchschnittliche Contiggröße ist 66.062 bp mit einer N50-Contiggröße[2] von 243.705 bp.

[2]Definition: Die Maßzahl N50 entspricht der Länge des kleinsten Contigs in einem Set der größten Contigs, deren zusammengesetzte Länge mind. 50 % des Assemblies entspricht (Thallinger 2011)

Solexa. Die „*paired-end*“ Solexa-Sequenzierung ergibt 1.229.929 Reads, die gegen das Genom N315 gemapped werden und dieses zu 99 % abdecken.

Die Nutzung von zwei Sequenziertechnologien hat den Vorteil, dass die geschlossene Sequenz mit einer sehr guten Qualität vorliegt und Sequenzierfehler nahezu ausgeschlossen werden können. Weiterführende Analysen unter Berücksichtigung der Genome N315, JH1 und JH9 des klonalen Komplexes CC5 ergaben 122 für 04-02981 spezifische Unterschiede (110 Basenaustausche, 2 Insertionen, 10 Deletionen), von denen lediglich sechs zwischen 454 und Solexa widersprüchlich waren: vier dieser Konflikte traten in Homopolymeren auf, die dafür bekannt sind, von 454 schlecht aufgelöst werden zu können (Stangier, Bauser & Regenbogen 2007, Shendure & Ji 2008).
Insgesamt ist die Fehlerrate von einem Fehler pro einer Million Basenpaare sehr gering und beide Methoden sind gut für die Rekonstruktion von Phylogenien auf genomischer Ebene geeignet.

5.1.2 Genometrische Daten

Allgemeine Merkmale des Genoms 04-02981 im Vergleich zu den Genomen N315 und JH1 sind in Tabelle 5.1 zusammengefasst. Die Genome sind sich in den untersuchten Merkmalen sehr ähnlich. Vergleichende Genomanalysen haben außerdem ein hohes Maß an Syntenie ergeben, was auch schon in früheren Analysen gezeigt wurde (Lindsay & Holden 2004). Die Sequenzidentität zu N315 bzw. JH1 beträgt 96 % bzw. 99 %; Unterschiede sind auf mobile genetische Elemente zurückzuführen. Ausführlichere Informationen zu den enthaltenen mobilen Elementen sind in Tabelle 5.5 in Kapitel 5.2.5 zusammengefasst.

Tabelle 5.1: Genetische Merkmale des Genoms 04-02981.

	04-02981	N315	JH1
Länge der Sequenz [bp]	2.821.452	2.813.641	2.906.507
G+C Gehalt			
Total	33	33	33
CDS	34	34	34
RNAs	51	50	52
intergenisch	31	28	30
CDS			
Total	2.732	2.595	2.747
Anteil im Genom [%]	83,8	84,5	83,7
Ribosomale RNAs			
16S	5	5	6
23S	5	5	6
5 S	7	6	7
Transfer RNAs	63	62	60
Transfer-Messenger RNAs	1	1	1

5.2 Mikroevolution des klonalen Komplexes CC5

5.2.1 Verwendete Isolate

Um die Phylogenie des klonalen Komplexes CC5 zu rekonstruieren, werden die zu sequenzierenden Isolate anhand der von Nübel *et al.* (2008) rekonstruierten Populationsstruktur des Sequenztyps ST5 ausgewählt (Abbildung 5.2). So wird gewährleistet, dass der Datensatz repräsentativ für die Populationsstruktur ist und jede phylogenetische Linie abgedeckt wird (rote Linien). Der Datensatz wurde um 14 weitere in GenBank vorhandene Stämme ergänzt.
Tabelle 5.2 enthält eine Übersicht der verwendeten Genome des klonalen Komplexes CC5 sowie zugehörige epidemiologische Daten und die verwendete Sequenziermethode.

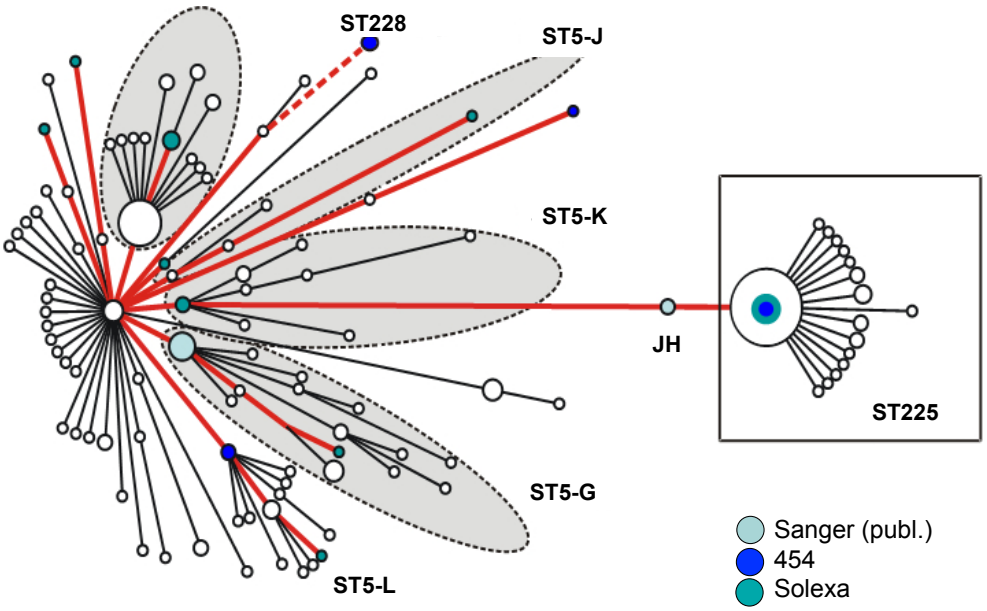

Abbildung 5.2: „*Guiding tree*“ zur Auswahl repräsentativer Isolate. Der „*Minimum Spanning Tree*“ zeigt die Populationsstruktur der Sequenztypen ST5 und ST225. Die roten Linien markieren die durch der Datensatz abgedeckten phylogenetischen Linien. Die farbigen Kreise repräsentieren die verwendeten Sequenziermethoden. (Quelle: nach Nübel *et al.* 2008 und Nübel *et al.* 2010)

Tabelle 5.2: Übersicht der verwendeten CC5 Genome.

Isolat (RKI-ID)	Resistenztyp	Sequenztyp	Herkunftsland	Datum der Isolation	Sequenzier-methode	Referenz
02-02424	MRSA	ST5	Deutschland	2002	Solexa	diese Arbeit
02-03179	MRSA	ST228	Deutschland	2002 (?)	454	diese Arbeit
04-02232	MRSA	n.a.	Deutschland	2004	Solexa	diese Arbeit
04-02981	MRSA	ST225	Deutschland	2004	454/Solexa	CP001844
05-00586	MRSA	n.a.	Deutschland	2006	454	diese Arbeit
06-00603	CA-MRSA	ST5	Deutschland	2006	454	diese Arbeit
06-01243	MRSA	ST5	Deutschland	2006	Solexa	diese Arbeit
07-00010	MRSA	ST5	USA	n.a.	Solexa	diese Arbeit
07-00170	MRSA	ST5	Südkorea	2004	Solexa	diese Arbeit
07-02020	MSSA	ST5	Südafrika	1996	Solexa	diese Arbeit
A5937	MRSA	ST5	USA	2000	454	ACKF00000000
A6224	MRSA.	ST5	USA	2001	454	ACKC00000000
A6300	MRSA	ST5	USA	2000	454	ACKE01000000
A8115	MRSA	ST5	USA	1997 (?)	454	ACKG01000000
A9299	MRSA	ST5	USA	2005	454	ACKH01000000
A9719	MRSA	ST5	USA	2004	454	ACKJ00000000
A9763	MRSA	ST5.	USA	2006	454	ACKK00000000
A9781	MRSA	ST5	USA	2006	454	ACKL01000000
ED98	Hühner-MRSA	Hühner-ST	UK	1997/1998	454	CP001781
JH1	MRSA	ST105	USA	2000	Sanger	CP000736
JH9	MRSA	ST105	USA	2000	Sanger	CP000703
MR1	MRSA	ST5	Polen	1992	Solexa	ACZQ00000000
Mu50	VISA	ST5	Japan	1997	Sanger	BA000017
N315	VSSA	ST5	Japan	1982	Sanger	BA000018

n.a. - Information nicht verfügbar

5.2.2 Sequenzierung und *de novo* Assemblierung bzw. Readmapping

Der Datensatz zur Analyse des klonalen Komplexes 5 umfasst insgesamt 24 Isolate. Die Genome von zehn Stämmen wurden extra für diese Arbeit mit den NGS-Technologien 454 und Solexa sequenziert. Tabelle 5.3 enthält eine Übersicht zu einigen Ergebnissen der Sequenzierungen sowie zur anschließenden Readauswertung. Die Ergebnisse der Sequenzierung des Genoms 04-02981 wurden bereits in Kapitel 5.1 vorgestellt.

Tabelle 5.3: Statistiken zur Sequenzierung der verwendeten Isolate.

454							
Isolat	**Anzahl Reads**	**Anzahl sequenzierter Basen**	**⌀ Read Länge [bp]**	**Anzahl Contigs > 500 bp**	**Länge aller > 500 bp Contigs**	**N50 Contiggröße**	**⌀ Coverage**
02-03179	461.685	121.030.435	262	79	2.831.678	99.087	n.a.
05-00586	313.178	n.a.	354	50	2.797.668	n.a.	n.a.
06-00603	370.769	98.613.613	266	74	2.835.219	63.614	26
Solexa							
Isolat	**Anzahl Reads**	**Anzahl sequenzierter Basen**	**⌀ Read Länge [bp]**	**Anzahl alignierter Reads**			**⌀ Coverage**
02-02424 *	2.919.356	105.096.816	36	2.369.880			30
04-02232	1.926.191	61.638.112	32	3.724.125			n.a.
06-01243	2.201.583	70.450.656	32	4.163.887			n.a.
07-00010	3.289.342	105.258.944	32	5.470.769			n.a.
07-00170 *	4.247.399	152.906.364	36	3.137.053			40
07-02020 *	4.710.286	169.570.296	36	n.a.			n.a.

n.a. - Information nicht verfügbar. * - „*paired-end*“ Sequenzierung

5.2.3 Die Phylogenie des klonalen Komplexes CC5

Zur Rekonstruktion der Phylogenie des klonalen Komplexes CC5 werden 2.971 variable, Qualitäts-geprüfte Positionen des Kerngenoms plus 3.260 SNPs, die den klonalen Komplex CC5 von der Außengruppe unterscheiden, verwendet. Der Begriff Kerngenom ist für diese Arbeit so definiert, dass mobile genetische Elemente, Wiederholungssequenzen und Bereiche, die nicht in allen Genomen vorkommen, von phylogenetischen Analysen ausgeschlossen sind.

Abbildung 5.3 zeigt die phylogenetischen Bäume der „*Maximum Likelihood*“- und Bayes'schen-Analyse. Beide Bäume zeigen die gleiche Topologie und lediglich die Unterstützung der Knoten variiert leicht. Es sind deutlich zwei Cluster zu sehen, die recht divers zueinander sind: Das erste umfasst den Großteil der Isolate mit unterschiedlicher geographischer Herkunft. Es enthält außerdem die bereits von Nübel *et al.* (2008) beschriebene ostasiatische Linie mit Isolaten aus Japan und Korea (Abbildung 1.4). Das zweite Cluster besteht aus fast allen US-amerikanischen sowie einigen deutschen Stämmen.

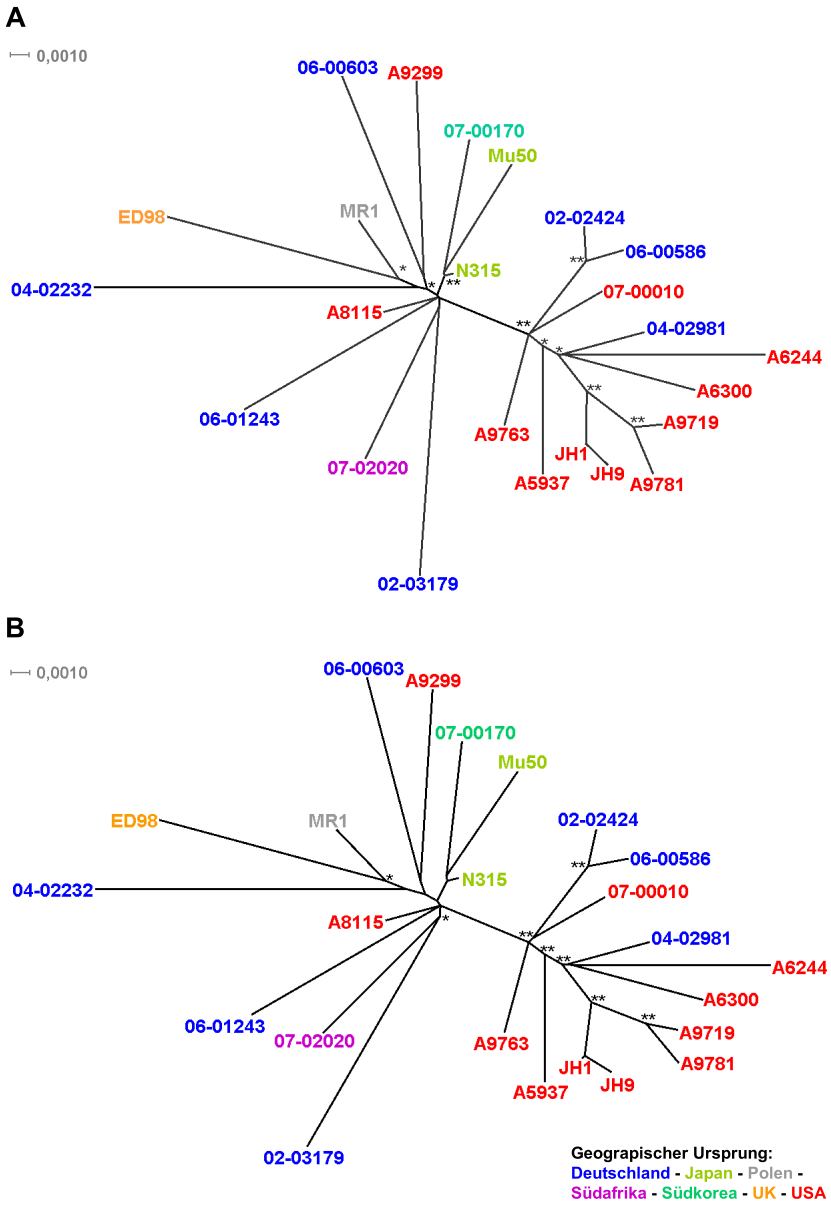

Abbildung 5.3: Phylogenie des klonalen Komplexes CC5 basierend auf 2.971 SNPs. A: „*Maximum Likelihood*“-Baum. B: Bayes'scher-Baum. ** approximierter Boostrap-Wert/„*posterior probability*“ = 100/1 * approximierter Boostrap-Wert/„*posterior probability*“ = 95 - 99/0,95 - 0,99. Der Maßstab zeigt jeweils die evolutionäre Distanz in Substitutionen pro Position.

5.2.4 Statistiken

Verteilung von SNPs entlang des Genoms. Wie in Abbildung 5.4 zu sehen, sind die SNPs des Kerngenoms gleichmäßig über das Genom verteilt und Regionen mit einer starken Akkumulation von Mutationen sind nicht vorhanden. Lediglich in den mobilen genetischen Elementen (z.B. Prophagen, Pathogenitätsinseln und SCC*mec*-Kassetten) treten, wie erwartet, SNPs gehäuft auf. Die durchschnittliche Anzahl von SNPs im Kerngenom liegt bei 333 pro Genom.

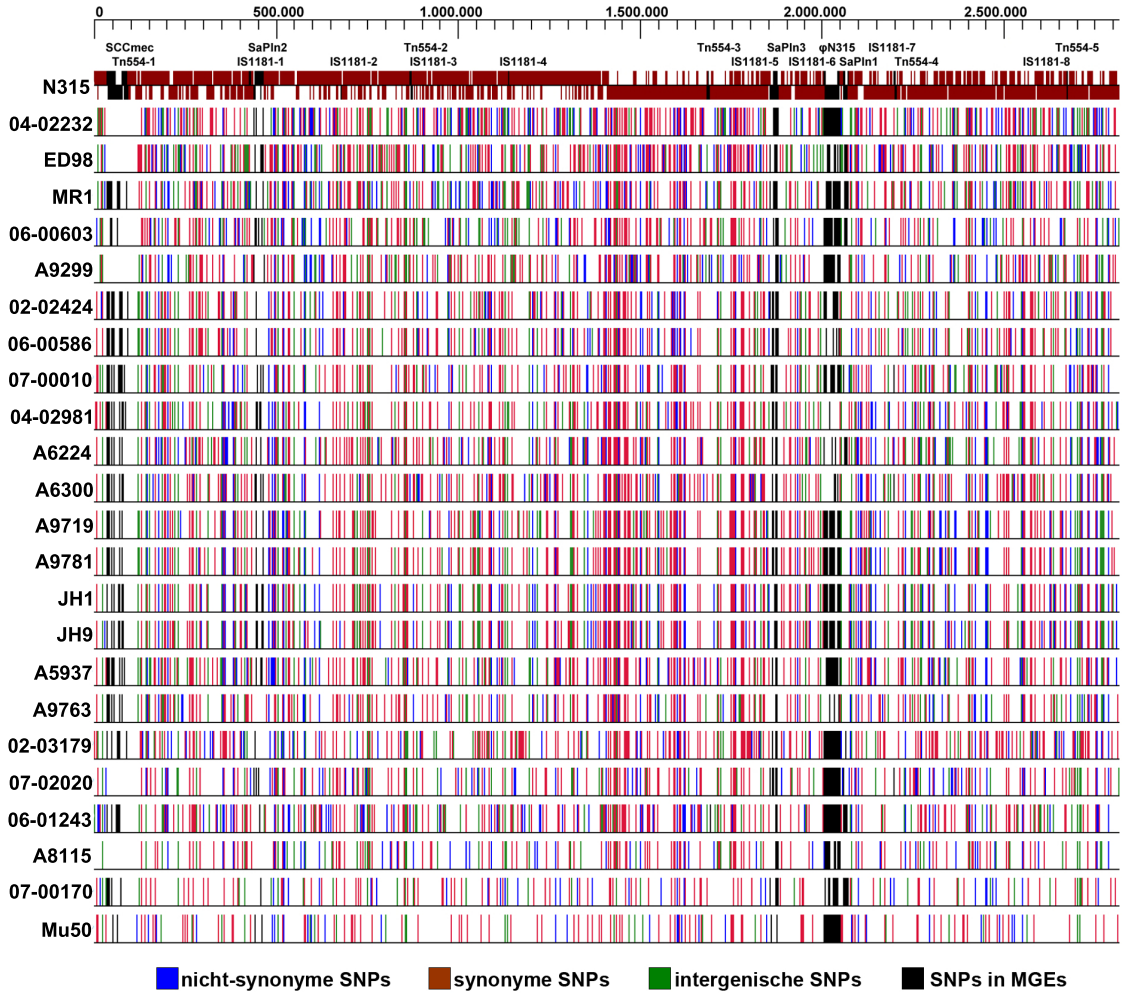

Abbildung 5.4: Verteilung von SNPs entlang des Genoms.

Verteilung von SNPs in funktionellen Genklassen. Neben einer gleichmäßigen Verteilung von SNPs über das Kerngenom, kann auch eine gleichmäßige Verteilung von Mutationen in verschiedenen funktionellen Genklassen beobachtet werden. Abbildung 5.5 zeigt für verschiedene Genklassen die Anzahl der SNPs im Kerngenom pro Nukleotid. Die Klassen sind gleichermaßen von Mutationen betroffen, wobei der Anteil nicht-synonymer SNPs gegenüber synonymen SNPs stark erhöht ist. Besonders groß ist allerdings die Zahl von SNPs in intergenischen Bereichen.

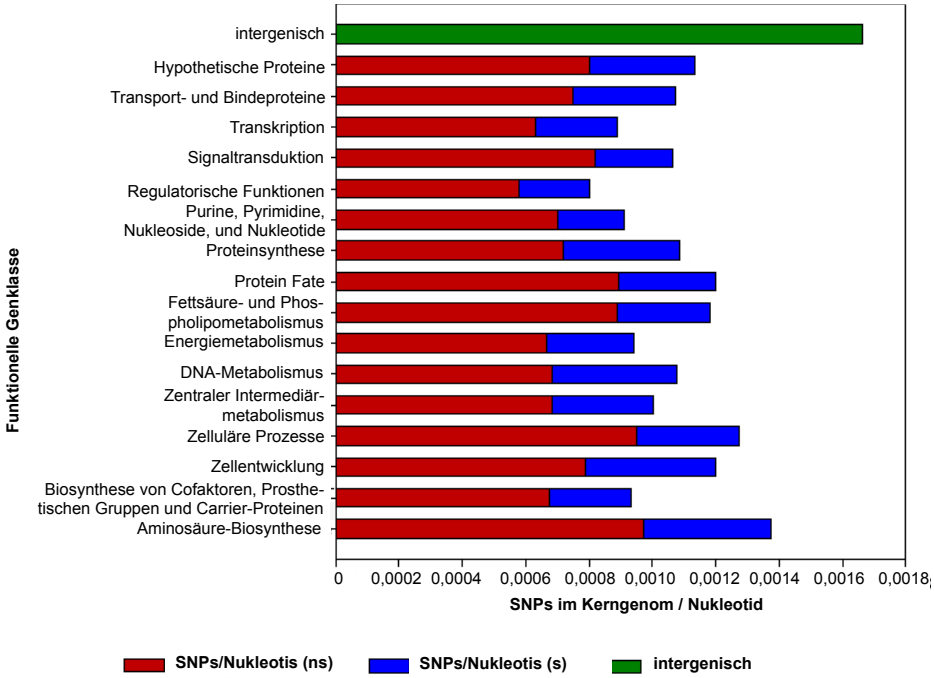

Abbildung 5.5: Verteilung von SNPs in funktionellen Genklassen.

Einfluss von Selektion: dN/dS. Das normalisierte Verhältnis nicht-synonymer (dN) zu synonymen (dS) Mutationen wird berechnet, um einen Hinweis auf den Einfluss von Selektion zu erhalten (Nei & Gojobori 1986). dN/dS wird paarweise für 24 CC5, zwei CC8 und ein CC30 Isolat ermittelt.

Vergleicht man sehr nah verwandte Spezies oder - wie in dieser Arbeit - Isolate einer Population, wird der evolutionäre Druck oft überschätzt, weshalb die Zeit seit der Diversifizierung berücksichtigt werden muss (Rocha *et al.* 2006). Aus diesem Grund wird dN/dS gegen die Anzahl intergenischer SNPs aufgetragen. Diese gelten als selektiv neutral und können somit als Ersatz für eine Zeiteinheit verwendet werden. Abbildung 5.6 zeigt dN/dS gegen die Anzahl intergenischer SNPs.

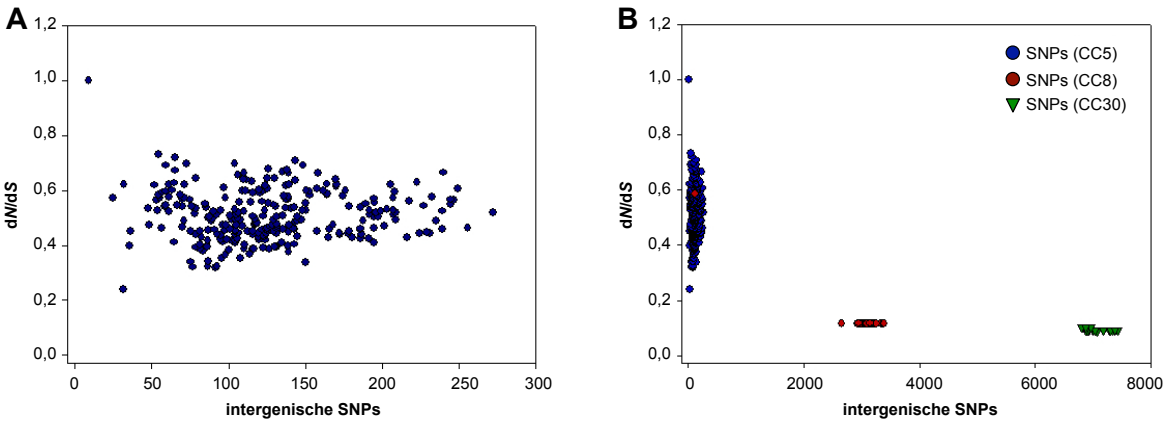

Abbildung 5.6: dN/dS gegen die Zeit. Paarweiser Vergleich zwischen A: 24 CC5 Isolaten. B: 24 CC5, zwei CC8 und einem CC30 Isolat.

Für den klonalen Komplex CC5 liegt der durchschnittliche dN/dS mit 0,51 unter 1, was auf einen mäßigen Einfluss reinigender Selektion hindeutet (Abbildung 5.6A). Zwischen den klonalen Komplexen dagegen ist dN/dS signifikant kleiner als 1 und der Einfluss reinigender Selektion damit stark ausgeprägt (Abbildung 5.6B).

Homoplasie. Unter den 2.971 SNPs sind lediglich sechs homoplastische SNPs (0,2 %), die alle durch Sanger-Sequenzierung bestätigt wurden (Tabelle 5.4). Der Homoplasie-Index, der als Maß zur Quantifizierung von Homoplasie dient, liegt bei 0,0041 (0: keine Homoplasie). Ein ähnlich geringer Anteil an Homoplasien wurde auch für *S. aureus* ST239 (Harris *et al.* 2010) und *Salmonella* Typhi (Holt *et al.* 2008) beschrieben.

Tabelle 5.4: Homoplasien in CC5.

Position in N315	Locus tag	Gen	Produkt	Aminosäure-Austausch
14603	intergenisch	-	-	-
229185	intergenisch	-	-	-
885644	intergenisch	-	-	-
2107968	SA1866	ilvA	Threonin Dehydratase	M→K
2240290	SA1972	-	Hypothetisches Protein, ähnelt einem „multidrug“-Transporter	H→Y
2308472	SA2048	rpsJ	30S ribosomales Protein S10	K→ M

5.2.5 Vergleichende Genomik

Das Kerngenom

Die Genomlänge der Isolate des klonalen Komplexes CC5 liegt bei ca. 2,8 Mb mit einer Gendichte von 0,94 Genen pro kb, was einem codierenden Bereich von 84 % entspricht. Der GC-Gehalt liegt bei 33 %.
Vergleichende Genomanalysen des klonalen Komplexes CC5 zeigen einen kolinearen Aufbau und ein stark konserviertes Kerngenom ohne Genaufnahme oder -verlust und stimmen damit mit bereits veröffentlichten Beschreibungen zwischen verschiedenen klonalen Komplexen überein (Lindsay & Holden 2004).

Mobile genetische Elemente

Bereiche, die nicht in allen Genomen vorkommen, werden als das akzessorische Genom bezeichnet. Aufgrund der hohen Konservierung der Genome des klonalen Komplexes CC5, besteht es vor allem aus mobilen genetischen Elementen, die horizontal zwischen verschiedenen Stämmen übertragen werden können. Zu diesen Elementen zählen Prophagen, Pathogenitätsinseln, genomische Inseln, Plasmide, Transposons, Insertionselemente und die chromosomale Kassette SCC*mec*. Einige dieser Elemente verbreiten sich

mit hoher Frequenz zwischen den Isolaten, wogegen andere wenig oder gar nicht übertragen werden (Lindsay & Holden 2004). Insgesamt machen mobile Elemente etwa 8 % der untersuchten Genome aus und das akzessorische Genom ist damit kleiner als von Lindsay & Holden (2004) beschrieben (25 %).
Tabelle 5.5 am Ende des Kapitels enthält eine Übersicht über die im klonalen Komplex CC5 vorkommenden mobilen genetischen Elementen.

SCC*mec*. Von den zur Zeit bekannten zwölf SCC*mec*-Typen, kommen die Typen I bis V in den vorliegenden Daten vor.

Pathogenitätsinseln. Die untersuchten Isolate tragen ein bis drei Pathogenitätsinseln (SaPIs).

SaPI1. Die Integrationsstelle der SaPI1 ist downstream des *groEL*-Gens lokalisiert. Sie enthält Gene, die die Enterotoxine L (*sel*) und C3 (*sec3*) sowie das „*toxic shock syndrome*"-Toxin-1 (*tst-1*) codieren. Vollständig ist SaPI1 nur in der ostasiatischen Klade vorhanden. Das Isolat 06-01243 enthält ebenfalls die drei Virulenzgene, wobei die Pathogenitätsinsel aber nicht komplett ist. Außerdem ist eine starke Anhäufung von Mutationen in der 06-01243 SaPI zu sehen, was sich deutlich von der Konservierung innerhalb der ostasiatischen Linie unterscheidet. Bei 06-01243 handelt es sich um ein mit Solexa sequenziertes Isolat. Aus diesem Grund können keine genaueren Aussagen über eine mögliche neue bzw. andere Insel getroffen werden.

SaPI2. SaPI2 integriert upstream des Gens *guaA* und kommt in allen analysierten Isolaten vor. Sie enthält ein Exotoxin-Gencluster (*set*) und ein Lipoprotein-Gencluster (*lpl*). *set* codiert einer Superantigen-Familie zugehörige Proteine, die die proinflammatorische Produktion von Zytokinen induziert (Kuroda *et al.* 2001). Das Cluster besteht aus zehn *set*-Homologen; eine Ausnahme bildet Mu50, in dem das Cluster nur neun homologene Gene umfasst. Das Lipoprotein-Cluster besteht in 22 Isolaten aus neun *lpl*-Homologen. In den Stämmen JH1 und JH9 ist ein *lpl*-Gen trunkiert. Insgesamt ist SaPI2 im klonalen Komplex CC5 stark konserviert und es sind kaum Indels und SNPs enthalten.

SaPI3. SaPI3 integriert downstream eines tRNA-Clusters und zwischen den hypothetischen Genen SA1621 und SA1649 (Locus Tag in N315). Die Pathogenitätsinsel enthält ein Serin Protease-Cluster (*spl*) und ein Enterotoxin-Cluster (*egc*) sowie die zwei Komponenten des Leukozidin DE Toxins (*lukDE*). Das Serin Protease Cluster besteht aus vier bis fünf paralogen Genen, die sekretorische Serin-Proteasen kodieren (Kuroda *et al.* 2001). Die Isolate 02-02424, 07-00170 und 07-02020 weisen

einige Trunkierungen in den Genen auf, die aber am Mapping liegen könnten. Das Enterotoxin-Cluster besteht aus den fünf Genen *seo*, *sem*, *sei*, *sek* und *seg* sowie den beiden Pseudogenen *yent1* und *yent2*, die homologe Enterotoxine kodieren. Alle Isolate enthalten dieses Cluster mit der gleichen Anzahl an Genen. Die Gene *lukD* und *lukE* kommen ebenfalls in allen untersuchten Isolaten vor. Insgesamt ist SaPI3 im klonalen Komplex CC5 stark konserviert.

ANDERE. Die Isolate 06-00603 und A6224 enthalten ein etwa 14.000 bp großes Stück, das zwischen den hypothetischen Genen SA0771 und SA0772 (Locus Tag in N315) integriert ist. Neben einigen Genen, die Genen in SaPI1 ähneln, kommen auch Bakteriophagen-Gene vor. Ein vorkommender Virulenzfaktor ist das Gen *ear*, welches für ein putatives β-Lactamase Protein codiert. Blast-Analysen der Region haben eine große Ähnlichkeit zu einem Genomabschnitts des Stamms *S. aureus* USA300-TCH60 (Akzessionsnummer: CP002110) ergeben. Zur Zeit kann keine Aussage über die genaue Art und phänotypische Auswirkungen dieses mobilen genetischen Elements gemacht werden.

Transposons und Insertionselemente. Das Transposon *Tn554* liegt in den untersuchten Isolaten bis zu fünfmal im Genom vor. *Tn554* enthält das Gen *spc*, das für eine Spectinomycin-Resistenz codiert. Eine spezielle Variante integriert innerhalb der SCC*mec*-Kassette und codiert zusätzlich eine Erythromycin-Resistenz (Kuroda *et al.* 2001). Die Isolate 02-02424, 04-02232, 06-00603, 06-01243, 07-00170, 07-02020, A8115 und A9299 besitzen gar keine *Tn554*-Transposasen. Mu50 und 07-00170 besitzen zusätzlich das Transposon *Tn5801* mit einer Tetrazyklin-Resistenz, die durch das Gen *tetM* codiert wird.
Das Insertionselement *IS1181* liegt mit bis zu acht Kopien im klonalen Komplex CC5 vor, was auch schon für andere *S. aureus* Stämme beschrieben wurde (Kuroda *et al.* 2001). Die Isolate 02-02424, 06-00586, 07-00170 und 07-02020 besitzen keine *IS1181*-Elemente.

Plasmide. Eine genaue Analyse über die Anzahl von Plasmiden und Plasmid-kodierte Pathogenitäts- und Virulenzeigenschaften wurde nicht durchgeführt. Jedoch kann eine Aussage über die An- bzw. Abwesenheit von Plasmiden in den untersuchten Isolaten getroffen werden (Tabelle 5.5). Demnach besitzen die Isolate 04-02981, 06-01243, 07-00010, A9299 und A9763 keine Plasmide.

Prophagen. Prophagen sind das im klonalen Komplex CC5 häufigste und diverseste mobile genetische Element. Die Tabellen 5.5 und 5.6 enthalten eine Übersicht über die in den Isolaten gefundenen Prophagen. Aufgrund ihrer Bedeutung für die Diversität

von *S. aureus*, werden die Prophagen im Kapitel 5.3 näher beschrieben.

Tabelle 5.5: Zusammenfassung der mobilen genetischen Elemente in CC5.*

	02-02424	**02-03179**	**04-02232**	**04-02981**	**06-00586**	**06-00603**	**06-01243**	**07-00010**	**07-00170**
SCC*mec*									
Typ I	-	-	-	-	-	-	✓	-	-
Typ II	✓	-	-	✓	✓	-	-	✓	-
Typ III	-	-	-	-	-	-	-	-	-
Typ IV	-	-	✓	-	-	-	✓	-	✓
Typ V	-	-	-	-	-	✓	-	-	-
Typ VI	-	-	-	-	-	-	-	-	-
Unbekannt	-	✓	-	-	-	-	-	-	-
Pathogenitätsinseln									
SaPI1	-	-	-	-	-	-	*sel*, *sec3*, *tst*(nicht vollständig)	-	*sel*, *sec3*, *tst*
SaPI2	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]
SaPI3	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*
andere	-	-	-	-	-	*ear*, ähnlich SaPI1, Integration zwischen SA0771 und SA0772	-	-	-
Transposons									
Tn554	-	*ermA* & *spc*[1]	-	*ermA* & *spc*[2]	*ermA* & *spc*[1]	-	-	*ermA* & *spc*[4]	-
Tn5801	-	-	-	-	-	-	-	*tetM*	
IS1181	-	*tnp*[1]	*tnp*[8]	*tnp*[6]	-	*tnp*[1]	*tnp*[8]	*tnp*[8]	-
Plasmide	✓	✓	✓	-	✓	✓	-	-	✓
Prophagen									
Sa1int	-	-	-	✓	✓	-	-	✓	-
Sa2int	-	-	-	-	-	✓	-	✓	✓
Sa3int	*chp*, *sak*, *scn*, *sep*	*chp*, *sak*, *scn*, *sep*	*sak*	*chp*, *sak*, *scn*	*chp*, *sak*, *scn*, *sep*	*sak*, *scn*	*chp*, *sak*, *scn*	*chp*, *sak*, *scn*	*chp*, *sak*, *scn*, *sep*
Sa4int	-	-	-	-	-	-	-	-	-
Sa5int	✓	-	-	-	-	-	-	-	✓
Sa6int	-	-	-	-	-	✓	-	-	-
Sa7int	✓	✓	-	-	-	-	-	-	✓

	07-02020	**A5937**	**A6224**	**A6300**	**A8115**	**A9299**	**A9719**	**A9763**	**A9781**
SCC*mec*									
Typ I	-	-	-	-	-	-	-	-	-
Typ II	-	✓	✓	✓	-	-	✓	✓	✓
Typ III	-	-	-	-	-	-	-	-	-
Typ IV	-	-	-	-	-	-	-	-	-
Typ V	-	-	-	-	-	-	-	-	-
Typ VI	-	-	-	-	-	-	-	-	-
Unbekannt	MSSA	-	-	-	✓	✓	-	-	-
Pathogenitätsinseln									
SaPI1	-	-	-	-	-	-	-	-	-
SaPI2	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]
SaPI3	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[4], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[4], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*
andere	-	-	*ear*, ähnlich SaPI1, Integration zwischen SA0771 und SA0772	-	-	-	-	-	-
Transposons									
Tn554	-	*ermA* & *spc*[1]	*ermA* & *spc*[1]	*ermA* & *spc*[1]	-	-	*ermA* & *spc*[1]	*ermA* & *spc*[1]	*ermA* & *spc*[1]
Tn5801	-	-	-	-	-	-			
IS1181	-	*tnp*[1]	*tnp*[1]	*tnp*[1]	*tnp*[1]	*tnp*[1]	*tnp*[1]	*tnp*[1]	*tnp*[1]
Plasmide	✓	✓	✓	✓	✓	-	✓	-	✓
Prophagen									
Sa1int	✓	✓	✓	-	✓	-	✓	✓	✓
Sa2int	-	✓	✓	✓	-	✓	-	-	-
Sa3int	*sak*, *scn*, *sep*	*chp*, *sak*, *scn*, *sep*	*chp*, *sak*, *scn*, *sep*	-	*chp*, *sak*, *scn*	*sak*, *scn*, *sep*	*chp*, *sak*, *scn*	*chp*, *sak*, *scn*, *sep*	*chp*, *sak*, *scn*
Sa4int	-	-	-	-	-	-	-	-	-
Sa5int	-	-	-	-	-	-	✓	-	✓
Sa6int	-	-	✓	-	-	-	-	-	-
Sa7int	-	-	-	-	-	-	-	-	-

* Angegeben sind Gene, die in Virulenz oder Antibiotika-Resistenz involviert sind. Abkürzungen: *chp* - Chemotaxis Inhibitor Protein, *ear* - putatives β–Lactamase Protein, *egc* - Enterotoxin-Cluster, *ermA* - Erythromycin-Resistenz, *lpl* - Lipoproteine, *lukDE* - Komponenten des Leukozidin DE Toxins, *sak* - Staphylokinase, *scn* - Staphylokokken Komplement Inhibitor, *sec3* - Enterotoxin C3, *sel* - Enterotoxin L, *sep* - Enterotoxin P, *set* - Staphylokokken Exotoxine, *spc* - Spectinomycin-Resistenz, *spl* - Staphylokokken Serinprotease, *tst* - Toxin des Toxik-Schock-Syndroms, *tetM* - Tetracyclin-Resistenz, *tnp* - Transposase
[x] Anzahl homologer Gene

Tabelle 5.5: Zusammenfassung der mobilen genetischen Elemente in CC5.*

	ED98	JH1	JH9	MR1	Mu50	N315
SCC*mec*						
Typ I	-	-	-	-	-	-
Typ II	-	✓	✓	-	✓	✓
Typ III	-	-	-	-	-	-
Typ IV	-	-	-	✓	-	-
Typ V	-	-	-	-	-	-
Typ VI	-	-	-	-	-	-
Unbekannt	✓	-	-	-	-	-
Pathogenitätsinseln						
SaPI1	-	-	-	-	*sel*, *sec3*, *tst*	*sel*, *sec3*, *tst*
SaPI2	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	*set*[10], *lpl*[9]	set [9] , lpl[9]	*set*[10] , *lpl*[9]
SaPI3	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[4], *lukDE*, *egc*	*spl*[4], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*	*spl*[5], *lukDE*, *egc*
andere	SaPIAv	-	-	ähnlich SaPI1 und SaPIAv	*tetM*	-
Transposons						
Tn554	-	*ermA* & *spc*[2]	*ermA* & *spc*[2]		*ermA* & *spc*[2]	*ermA* & *spc*[5]
Tn5801	-	-	-	-	*tetM*	-
IS1181	*tnp*[8]	*tnp*[8]	*tnp*[8]	*tnp*[8]	*tnp*[8]	*tnp*[8]
Plasmide	✓	✓	✓	✓	✓	✓
Prophagen						
Sa1int	✓	✓	✓	-	✓	-
Sa2int	-	-	-	✓	-	-
Sa3int	✓	*chp*, *sak*, *scn*	*chp*, *sak*, *scn*	*chp*, *sak*, *scn*	*sak*, *scn*, *sep*	*chp*, *sak*, *scn*, *sep*
Sa4int	-	✓	✓	-	-	-
Sa5int	-	-	-	-	-	-
Sa6int	-	✓	✓	-	-	-
Sa7int	-	-	-	-	-	-

* Angegeben sind Gene, die in Virulenz oder Antibiotika-Resistenz involviert sind. Abkürzungen: *chp* - Chemotaxis Inhibitor Protein, *ear* - putatives β–Lactamase Protein, *egc* - Enterotoxin-Cluster, *ermA* - Erythromycin-Resistenz, *lpl* - Lipoproteine, *lukDE* - Komponenten des Leukozidin DE Toxins, *sak* - Staphylokinase, *scn* - Staphylokokken Komplement Inhibitor, *sec3* - Enterotoxin C3, *sel* - Enterotoxin L, *sep* - Enterotoxin P, *set* - Staphylokokken Exotoxine, *spc* - Spectinomycin-Resistenz, *spl* - Staphylokokken Serinprotease, *tst* - Toxin des Toxik-Schock-Syndroms, *tetM* - Tetracyclin-Resistenz, *tnp* - Transposase
[x] - Anzahl homologer Gene

5.3 Prophagen im klonalen Komplex CC5

5.3.1 Genometrie

Jedes der 24 untersuchten *S. aureus* Isolate des klonalen Komplexes CC5 enthält ein bis vier Prophagen (Median: 2, Mittelwert: 1), mit einer Gesamtzahl von 58 Prophagen (Tabellen 5.5 und 5.6). Die Prophagen werden zur Familie der *Siphoviridae* gezählt und wurden nach Goerke *et al.* (2009) anhand ihres Integrase-Gens identifiziert.
Insgesamt wurden sieben Integrase-Gruppen gefunden, unter denen Sa1*int* (13x) und Sa3*int* (23x) am häufigsten vorkommen. Jede Integrase-Gruppe hat eine spezifische Insertionsstelle, die sowohl intergenisch als auch innerhalb von Genen liegen kann.
Die Genomgröße der Prophagen variiert zwischen 28 und 47 kb mit einem Median von 43 kb, was kongruent mit früheren Beschreibungen ist (Kwan *et al.* 2005, Lindsay 2008). Einige Prophagensequenzen sind allerdings deutlich kürzer. Dies kann damit erklärt werden, dass die Prophagen nicht komplett auf einem Contig lagen (bei Sequenzierung mit 454) bzw. Prophagen in Solexa-sequenzierten Stämmen schwer zu identifizieren waren (Kapitel 4.10).
Die Prophagen besitzen einen GC-Gehalt von 33 bis 36 % (Median: 33,5 %), der zwischen den verschiedenen Integrase-Gruppen leicht variiert und der dem *S. aureus* Genom sehr ähnlich ist. Diese Beobachtung wurde bereits für Mykobakteriophagen gemacht (Pedulla *et al.* 2003).

Der Gengehalt pro Prophage liegt bei 34 bis 89 Genen (Median: 71 Gene) und zeigt aus den bereits weiter oben beschriebenen Gründen, ebenfalls eine gewisse Diskrepanz. Die durchschnittliche Länge einer CDS umfasst 187 Aminosäuren (AA) und ist damit im Vergleich zu den Genen im Kerngenom des bakteriellen Wirts *S. aureus* viel kleiner (298 AA). Der Gesamtanteil kodierender Regionen ist allerdings mit einem Median von 96,4 % im Prophagen signifikant höher als in *S. aureus* CC5 (85 %).
Tabelle 5.6 enthält eine Übersicht der ermittelten Genometrie-Daten der Prophagen in den untersuchten Isolaten.

Tabelle 5.6: Übersicht zur Genometrie der 58 *S. aureus* CC5-Prophagen.

Nomenklatur	Länge [bp]	GC-Gehalt [%]	Integrase-Gruppe	MCL Cluster	Anzahl CDS	∅ CDS Länge [bp]	Anteil kodierender Bereiche [%]	Integrationsstelle
Sa1_04-02981	43623	34,4	Sa1*int*	I	73	586	98	SA0778 (*sufB*)/ SA0779
Sa1_07-00010	43920	34,55	Sa1*int*	I	89	478	96,9	n.a.
Sa1_07-02020	32790	35,73	Sa1*int*	I	47	681	97,7	n.a.
Sa1_A5937	43523	34,38	Sa1*int*	I	73	586	98,3	SA0778 (*sufB*)/ SA0779
Sa1_A6224	46564	33,47	Sa1*int*	I	78	541	90,6	SA0778 (*sufB*)/ SA0779
Sa1_A8115	37457	34,34	Sa1*int*	I	61	595	96,9	SA0778 (*sufB*)/ SA0779
Sa1_A9719	42520	34,59	Sa1*int*	I	69	604	98,1	SA0778 (*sufB*)/ SA0779
Sa1_A9763	43524	34,38	Sa1*int*	I	73	584	98,1	SA0778 (*sufB*)/ SA0779
Sa1_A9781	43066	34,36	Sa1*int*	I	74	572	98,4	SA0778 (*sufB*)/ SA0779
Sa1_ED98	42559	34,05	Sa1*int*	II	72	573	97	SA0778 (*sufB*)/ SA0779
Sa1_JH1	43623	34,43	Sa1*int*	I	59	645	87,2	SA0778 (*sufB*)/ SA0779
Sa1_JH9	43623	34,43	Sa1*int*	I	59	645	87,2	SA0778 (*sufB*)/ SA0779
Sa1_Mu50B	44079	34,05	Sa1*int*	I	70	572	90,7	SA0778 (*sufB*)/ SA0079
Sa2_06-00603	45791	33,41	Sa2*int*	IV	70	643	98,3	SA1319/SA1320
Sa2_07-00010	44685	33,39	Sa2*int*	IV	89	485	96,6	n.a.
Sa2_07-00170	44778	32,86	Sa2*int*	IV	79	554	97,8	n.a.
Sa2_A5937	44502	33,38	Sa2*int*	IV	70	620	97,5	SA1319/SA1320
Sa2_A6224	44530	33,36	Sa2*int*	IV	70	630	99	SA1319/SA1320
Sa2_A6300	41568	33,84	Sa2*int*	IV	49	774	91,3	SA1319/SA1320
Sa2_A9299	43830	33,32	Sa2*int*	IV	67	633	96,8	SA1319/SA1320
Sa2_MR1	42297	33,24	Sa2*int*	IV	90	464	98,8	n.a.
Sa3_02-02424	44560	32,87	Sa3*int*	II	104	415	97	n.a.
Sa3_02-03179	42226	33,16	Sa3*int*	II	76	538	96,8	SA1752 (*hlb*)
Sa3_04-02232	36466	33,99	Sa3*int*	II	83	435	99	n.a.
Sa3_04-02981	43800	32,84	Sa3*int*	II	76	553	96,1	SA1752 (*hlb*)
Sa3_06-00586	43798	32,81	Sa3*int*	II	78	540	96,2	SA1752 (*hlb*)
Sa3_06-00603	38720	33,29	Sa3*int*	II	70	541	97,8	SA1752 (*hlb*)
Sa3_06-01243	28158	33,53	Sa3*int*	II	77	362	99	n.a.
Sa3_07-00010	41299	33,48	Sa3*int*	II	85	460	94,8	n.a.
Sa3_07-00170	43433	32,74	Sa3*int*	II	89	465	95,4	n.a.
Sa3_07-02020	42555	33,11	Sa3*int*	II	76	543	96,9	n.a.
Sa3_A5937	43080	33,14	Sa3*int*	II	76	541	95,5	SA1752 (*hlb*)
Sa3_A6224	43264	32,83	Sa3*int*	II	77	540	96,1	SA1752 (*hlb*)
Sa3_A8115	37771	33,47	Sa3*int*	II	67	534	94,8	SA1752 (*hlb*)
Sa3_A9299	43372	33,32	Sa3*int*	II	67	633	96,8	SA1752 (*hlb*)
Sa3_A9719	41200	33,07	Sa3*int*	II	73	542	96	SA1752 (*hlb*)
Sa3_A9763	43578	32,86	Sa3*int*	II	76	549	95,8	SA1752 (*hlb*)
Sa3_A9781	41558	33,09	Sa3*int*	II	74	540	96,1	SA1752 (*hlb*)
Sa3_ED98	45905	32,85	Sa3*int*	IV	63	647	88,9	SA1752 (*hlb*)
Sa3_JH1	42159	33,07	Sa3*int*	II	67	572	90,9	SA1752 (*hlb*)
Sa3_JH9	42159	33,1	Sa3*int*	II	62	596	88,1	SA1752 (*hlb*)
Sa3_MR1	39682	32,94	Sa3*int*	II	69	544	94,6	n.a.
Sa3_Mu50A	42596	33,3	Sa3*int*	II	65	592	90,4	SA1752 (*hlb*)
Sa3_N315	43800	32,84	Sa3*int*	II	65	598	88,7	SA1752 (*hlb*)

n.a. - Information nicht verfügbar

Tabelle 5.6: Übersicht zur Genometrie der 58 *S. aureus* CC5-Prophagen.

Nomenklatur	Länge [bp]	GC-Gehalt [%]	Integrase-Gruppe	MCL Cluster	Anzahl CDS	∅ CDS Länge [bp]	Anteil kodierender Bereiche [%]	Integrationsstelle
Sa4_JH1	45151	33,66	Sa4*int*	IV	64	670	95	SA0878/SA0879
Sa4_JH9	45151	33,65	Sa4*int*	IV	61	695	94,5	SA0878/SA0879
Sa5_02-02424	34655	34,85	Sa5*int*	I	98	405	97,2	n.a.
Sa5_07-00170	34387	34,26	Sa5*int*	I	73	454	96,4	n.a.
Sa5_A9719	39498	34,18	Sa5*int*	I	57	664	95,8	SAS054/SA1693
Sa5_A9781	39553	34,17	Sa5*int*	I	59	644	96	SAS054/SA1693
Sa6_06-00603	31488	33,46	Sa6int	II	62	494	97,3	SA0309 (*geh*)
Sa6_A6224	30915	35,99	Sa6int	I	49	608	96,4	SA0309 (*geh*)
Sa6_JH1	42264	35,37	Sa6int	I	66	611	95,4	SA0309 (*geh*)
Sa6_JH9	42264	35,37	Sa6int	I	63	624	93,1	SA0309 (*geh*)
-_06-00586	29734	34,65	n.a.	I	46	641	99,2	SAS033/SA0976
Sa7_02-02424	32450	34,81	Sa7int	I	75	426	98,4	n.a.
Sa7_02-03179	44248	35,73	Sa7int	I	74	575	96,2	SAS033/SA0976
Sa7_07-00170	34451	34,31	Sa7int	I	77	432	96,6	n.a.
PH15 (*S. epidermis*, Siphoviridae)	44041	34,91	n.a.	Außengruppe	48	627	68,4	-
Mittelwert	41107	33,80			71,03	563,59	95,52	
Median	42578	33,48			71,00	572,00	96,40	
Min	28158	33			34	362	87	
Max	46564	36			89	774	99	

n.a. - Information nicht verfügbar

5.3.2 Klassifizierung von Prophagen

Wie in Kapitel 4.10.3 beschrieben, wurden zur Klassifizierung der Prophagen verschiedene Methoden verwendet. Da die erste Methode nach Goerke *et al.* (2009) nur auf dem Integrase-Gen basiert, wurden zwei weitere Ansätze durchgeführt.
Der erste beruht auf einem paarweisen Vergleich der gesamten Genomsequenz auf Nukleotidebene (Abbildung 5.7A). Insgesamt ergeben sich vier Cluster, die -mit wenigen Ausnahmen- die Integrase-Gruppen repräsentieren. Cluster 1 umfasst die Integrase-Gruppe Sa6*int*. Das zweite Cluster beinhaltet Sa1*int*, Sa5*int*, Sa6*int* und Sa7*int* sowie den Prophagen -_06-00586, dessen Sequenz nicht vollständig vorliegt und dem deswegen keine Integrase zugeordnet werden kann. Cluster 3 umfasst neben der großen Gruppe der Sa3*int*-Prophagen auch je ein Isolat von Sa1*int* (Sa1_ED98) und Sa6*int* (Sa6_06-00603). Cluster 4 setzt sich aus den Gruppen Sa2*int* und Sa4*int* zusammen und enthält den Ausreißer Sa3_ED98, der zur Sa3*int*-Gruppe gehört.
Die zweite Analyse beruht auf Protein-Familien (Abbildung 5.7B). Insgesamt wurden 4168 Proteine in 132 Familien zusammengefasst. Die vier Cluster, die auf Nukleotidebene ermittelt wurden, werden durch diese Analyse bestätigt. Lediglich innerhalb der Cluster gibt es leichte Unterschiede in der Topologie.

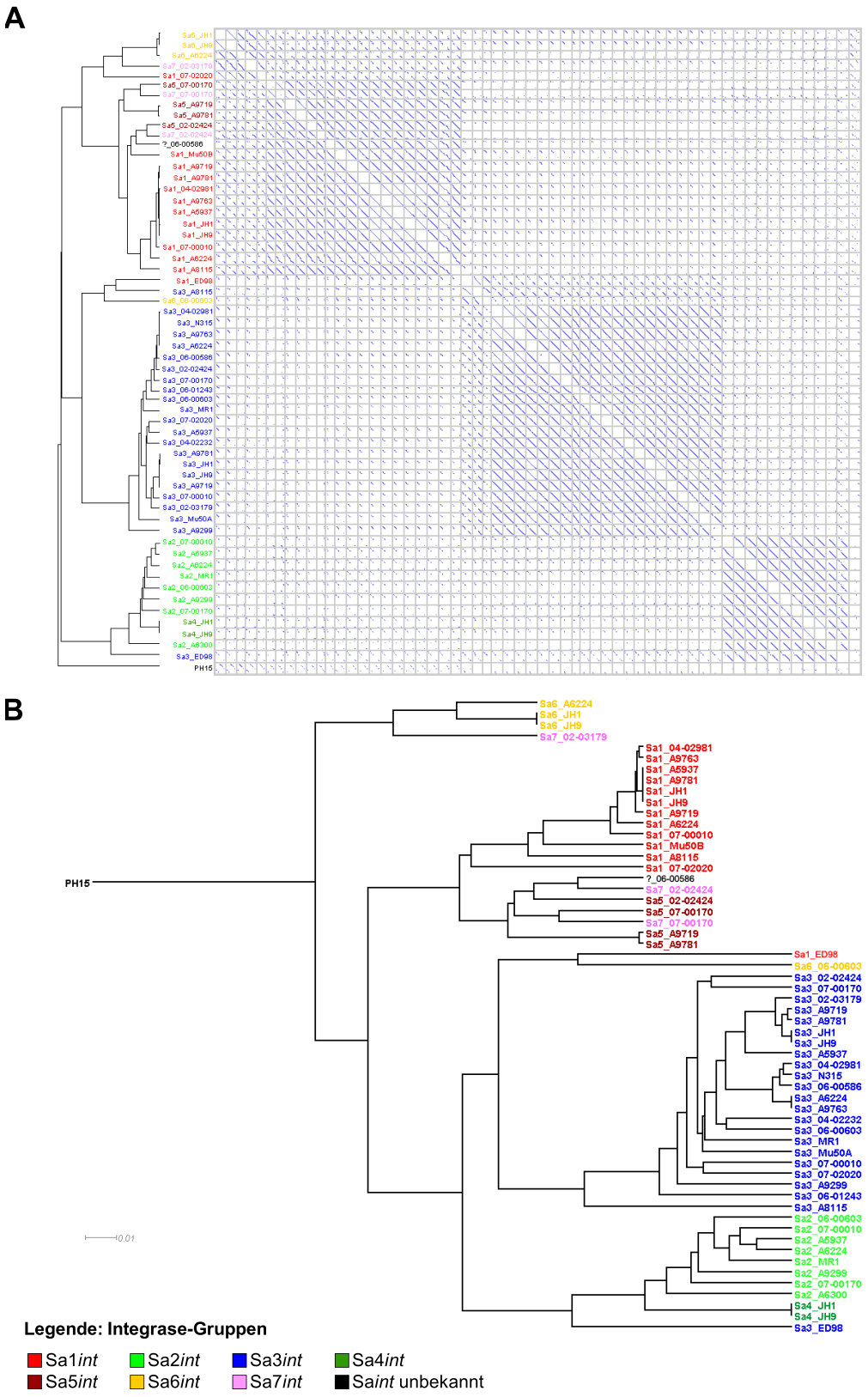

Abbildung 5.7: Cluster von *S. aureus* CC5 Prophagen basierend auf (A) Genom- und (B) Proteom-Vergleichen basierend auf UPGMA. A: Berechnung der Dotplot-Matrix mit Kodon. B: Gewurzelter UPGMA-Baum basierend auf CDS in den CC5-Prophagen. Sequenzen wurden mit MCL und einem E-Wert Grenzwert von 10^{-20} geclustert. Der Baum wurde mit PAUP* generiert.

5.3.3 Das Prophagen-Genom

Allgemeines. Die Genome der untersuchten Prophagen zeichnen sich durch eine ausgeprägte Konservierung in der Anordnung und Reihenfolge der Gene aus, was frühere Beschreibungen zur Organisation der *Siphoviridae* im Allgemeinen (Canchacha *et al.* 2009) und in *S. aureus* (Lindsay 2008) bestätigen.

Diversität. Prophagen des klonalen Komplexes CC5 weisen, verglichen mit der starken Konservierung im bakteriellen Wirtsgenom, ein hohes Maß an Diversität auf. Die Diversität reicht von einzelnen SNPs bis hin zu Regionen, die so gut wie keine Ähnlichkeit mehr zeigen, selbst wenn der Prophage zur gleichen Integrase-Gruppe gezählt wird (Abbildung 5.8).

Konversion. Die Integration von Prophagen in das Bakterien-Genom kann sowohl zu positiver als auch negativer Konversion führen. Zusätzliche Gene sind häufig Virulenzfaktoren und bringen dem bakteriellen Wirt Fitnessvorteile (positive Konversion). Analysen zur genetischen Ausstattung der untersuchten Prophagen des klonalen Komplexes CC5 haben ergeben, dass lediglich die Sa3*int*-Prophagen mit bekannten Pathogenitäts- und Virulenzgenen ausgestattet sind. Diese Gene sind *chp* (Chemotaxis Inhibitor Protein) *scn* (Staphylokokken Komplement Inhibitor), *sak* (Staphylokinase) und *sep* (Enterotoxin P).
Acht der 23 untersuchten Sa3*int*-Phagen besitzen alle vier Gene; die restlichen kodieren unterschiedliche Kombinationen. Die große Anzahl verschiedener Virulenzgen-Kombinationen, die Goerke *et al.* (2009) für Sa3*int* Phagen beobachtet haben, wird für die in dieser Arbeit untersuchten Phagengenome also bestätigt.
Allerdings muss darauf hingewiesen werden, dass aufgrund der schwierigen *de novo* Assemblierung von kurzen Solexa Reads keine endgültige Aussage ueber das Vorhanden sein bzw. die Abwesenheit von Genen gemacht werden kann.
Negative Konversion tritt auf, wenn ein Prophage in ein im Kerngenom-kodiertes Virulenzgen integriert und dieses so zerstört wird. Prophagen der Integrase-Gruppe Sa3*int* integrieren in das Gen *hlb*, welches das Toxin β-Hämolysin kodiert. β-Hämolysin ist sowohl in die Zellinvasion als auch an der Schädigung des Wirts durch die Auflösung von Erythrozyten involviert. Sa6*int*-Prophagen integrieren in das Lipase-kodierende Gen *geh*. Lipasen sind als Invasine ebenfalls an der Invasion in die Wirtszelle beteiligt. Aufgrund der großen Anzahl verschiedener Virulenzfaktoren in *S. aureus*, werden diese Aufgaben von anderen Faktoren übernommen.

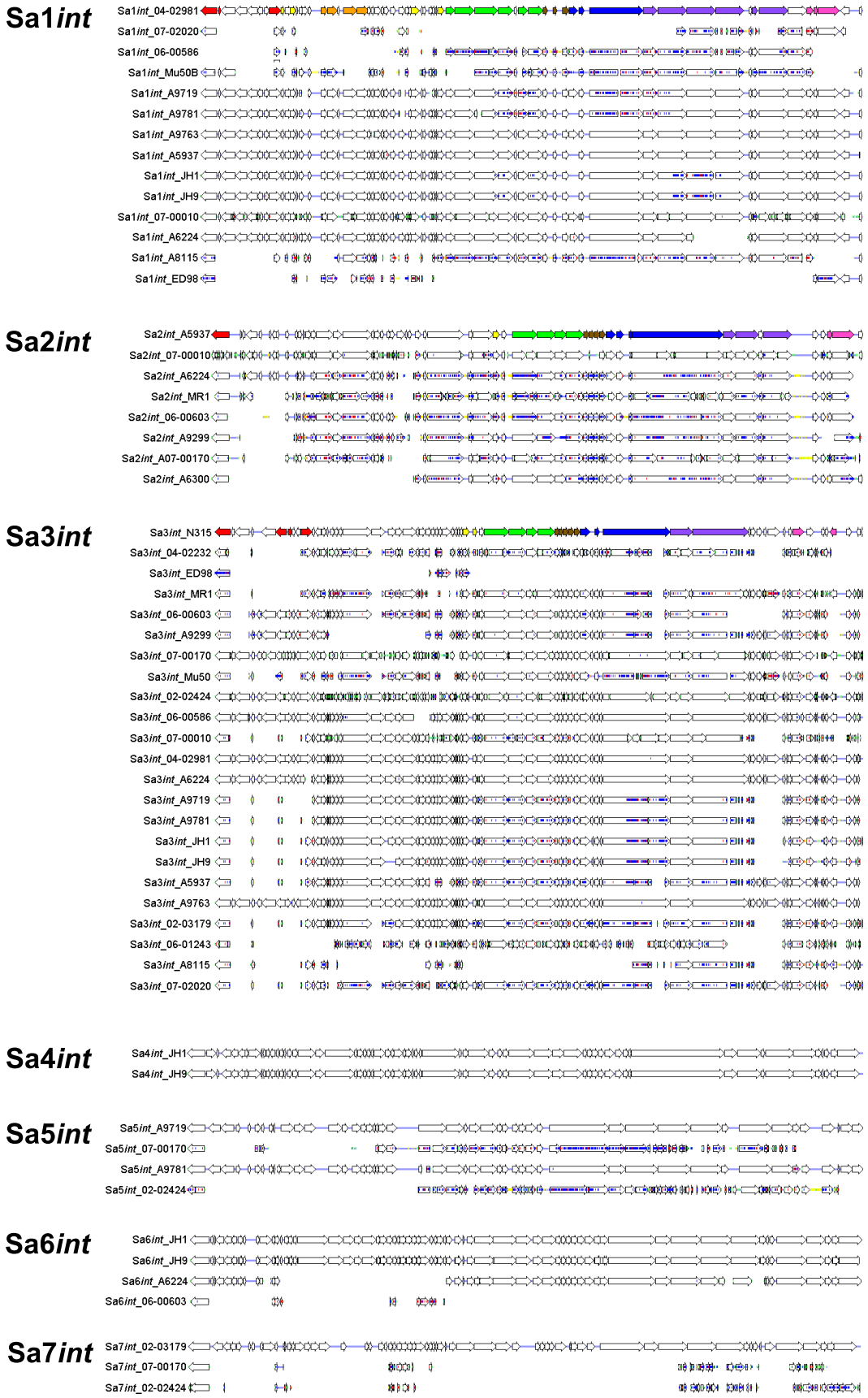

Abbildung 5.8: Diversität in Prophagen.

5.3.4 Rekombination in Prophagen

Die Grundlage zur Detektion von Rekombination in Prophagen mit ClonalFrame (Didelot & Falush 2007) ist ein gutes Alignment. Da zu große Variabilität, z.B. komplett unterschiedliche Sequenzabschnitte, schwer zu alignieren sind, wurden zu diverse Prophagen aus den Analysen ausgeschlossen. Außerdem wurden Sequenzabschnitte, die nicht in mindestens der Hälfte der verwendeten Sequenzen vorkommen, aus dem Alignment entfernt. So können robuste Untersuchungen durchgeführte werden.
Da die Rekombinationsereignisse im nächsten Schritt auf die Phylogenie der Bakterien abgebildet werden sollen (Kapitel 4.11.2), wurden nur Bakterien-Genome verwendet, deren Prophage auch im Datensatz enthalten ist.
Die Prophagen der Integrase-Gruppen Sa1*int*, Sa2*int* und Sa3*int* sind im klonalen Komplex CC5 am häufigsten vertreten. Da die verbleibenden drei Gruppen entweder zu divers sind oder die Anzahl der Prophagen zu gering ist, um Rekombination detektieren zu können, werden mit ihnen keine weitergehenden Analysen durchgeführt.

Rekombination in Sa1*int*. Innerhalb der Sa1*int*-Gruppe werden die vier Prophagen Sa1_07-02020, Sa1_ED98, Sa1_A8115 und Sa1_Mu50 aufgrund ihrer großen Diversität ausgeschlossen. Das Alignment der übrigen Sequenzen hat eine Länge von 43.654 bp mit 572 variablen Positionen (1 %). Das Abbilden der Rekombinationereignisse auf die Bakterien-Phylogenie (Abbildung 5.9A) ergibt 113 rekombinierte Fragmente. Die Länge der Fragmente variiert zwischen 2 und 2.901 bp mit einem Median von 61 bp.
Der Mantel-Test ergibt einen Korrelations-Koeffizenten r = 0,39 mit einer assoziierten Wahrscheinlichkeit von p = 0,037. Damit ist p auf dem 5 %-Niveau signifikant. Eine Korrelation zwischen der Rekombination in Sa1*int*-Prophagen und der Distanz der Wirts-Bakterien ist gegeben (Abbildung 5.9B).

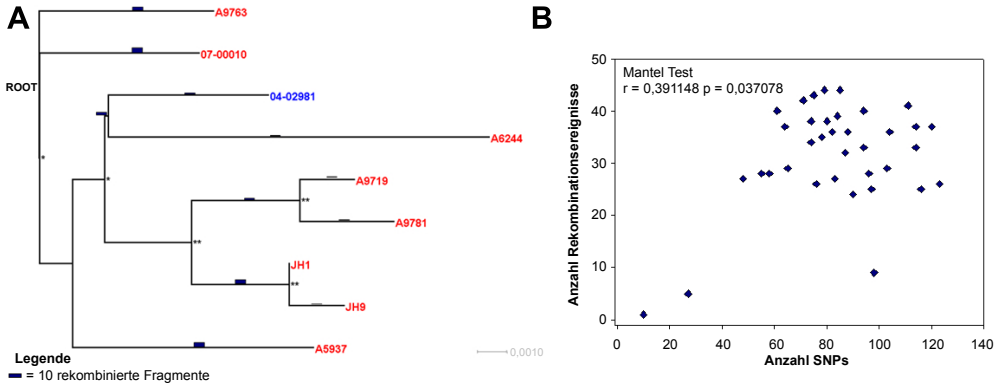

Abbildung 5.9: Rekombination in Sa1*int*. A: Abbildung der Rekombinationsereignisse auf die CC5-Phylogenie. B: Korrelation der paarweisen Bakterienendistanz mit der Anzahl der Rekombinationsereignisse in den Sa1*int*-Prophagen.

Rekombination in Sa2*int*. Für die Rekombinationsanalyse werden alle Sa2*int*-Prophagen verwendet. Sequenzabschnitte im Alignment, die nicht in mindestens vier der acht Sequenzen vorkommen, werden von der Analyse ausgeschlossen, wodurch sich eine Alignment der Länge 42.730 bp mit 2.825 variablen Positionen (7 %) ergibt. Die 356 gefundenen Rekombinationsereignisse haben eine Größe von 2 bis 38.481 bp mit einem Median von 325 bp und sind fast ausschließlich an den Spitzen des Wirts-Stammbaums lokalisiert (Abbildung 5.10A).
Der Mantel-Test ergibt einen Korrelations-Koeffizienten r = 1 mit einem Wahrscheinlichkeitswert p = 0,00005, womit Signifikanz zwischen der Wirtsdiversität und der Anzahl rekombinierter Fragmente in den Sa2*int*-Prophagen gegeben ist (Abbildung 5.10B).

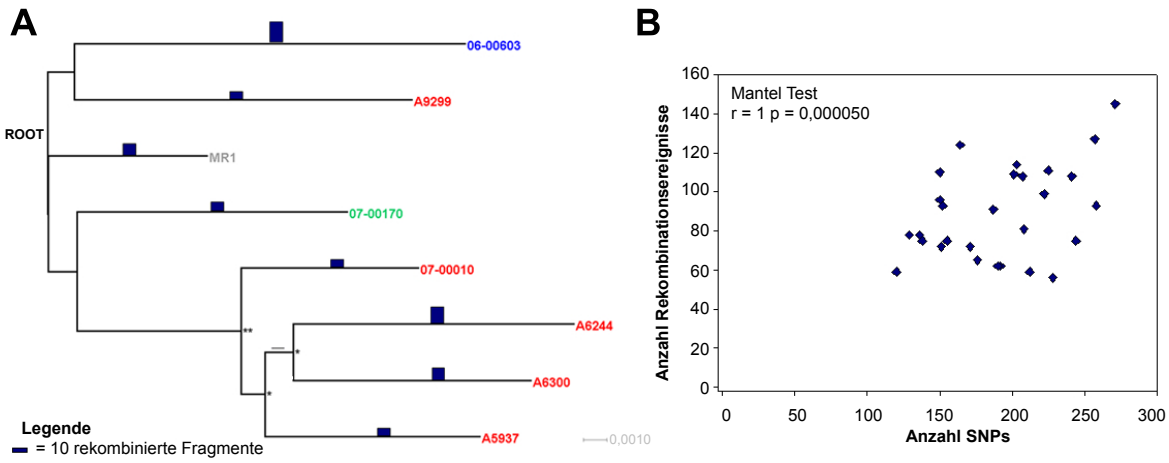

Abbildung 5.10: Rekombination in Sa2*int*. A: Abbildung der Rekombinationsereignisse auf die CC5-Phylogenie. B: Korrelation der paarweisen Bakterienendistanz mit der Anzahl der Rekombinationsereignisse in den Sa2*int*-Prophagen.

Rekombination in Sa3*int*. Innerhalb der Sa3*int*-Gruppe werden die vier Prophagen Sa3_ED98, Sa3_A9299, Sa3_A8115 und Sa3_Mu50 ausgeschlossen. Außerdem werden im Alignment nur Sequenzbereiche berücksichtigt, die in mindestens zehn der 19 verwendeten Prophagen-Sequenzen vorhanden sind. Das fertige Alignment hat eine Gesamtlänge von 39.989 bp mit 3.653 variablen Positionen (9 %). Das Mappen der Rekombinationsereignisse auf die Phylogenie (Abbildung 5.11A) ergibt 1.720 rekombinierte Fragmente. Die Größe der Fragmente liegt zwischen 2 und 1.288 bp mit einem Median von 13 bp. Abbildung 5.12 zeigt die Größenverteilung der Fragmente.
Der Mantel-Test ergibt einen Korrelations-Koeffizenten r = 0,36 mit einer assoziierten Wahrscheinlichkeit von p = 0,0011. Die Korrelation zwischen der Rekombination in Sa3*int*-Prophagen und der Distanz der Wirts-Bakterien ist signifikant (Abbildung 5.11B). Ab einer Diversität von ca. 130 SNPs zwischen den Wirtssequenzen ist eine Sättigung der Anzahl der Rekombinationsereignisse in den Prophagen zu erkennen.

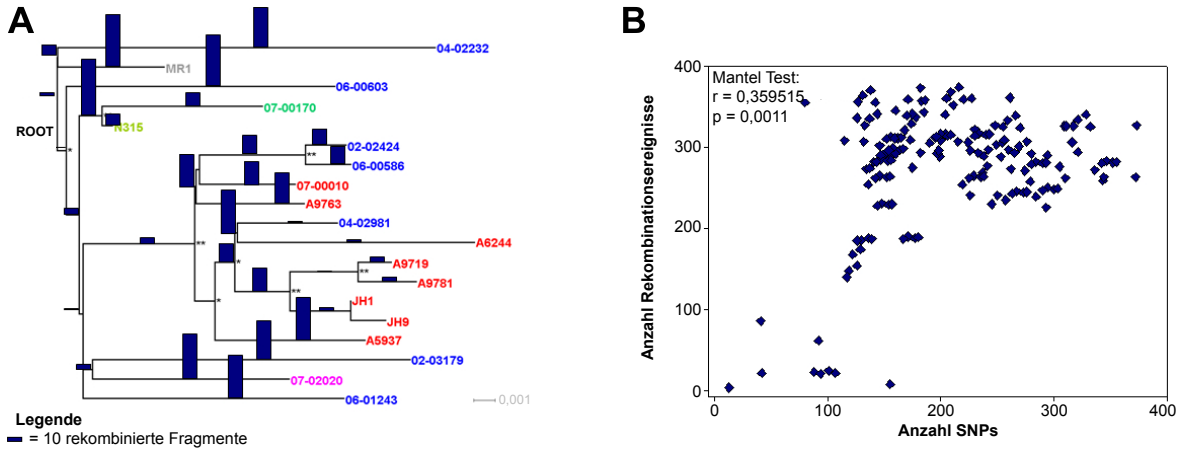

Abbildung 5.11: Rekombination in Sa3*int*. A: Abbildung der Rekombinationsereignisse auf die CC5-Phylogenie. B: Korrelation der paarweisen Bakterienendistanz mit der Anzahl der Rekombinationsereignisse in den Sa3*int*-Prophagen.

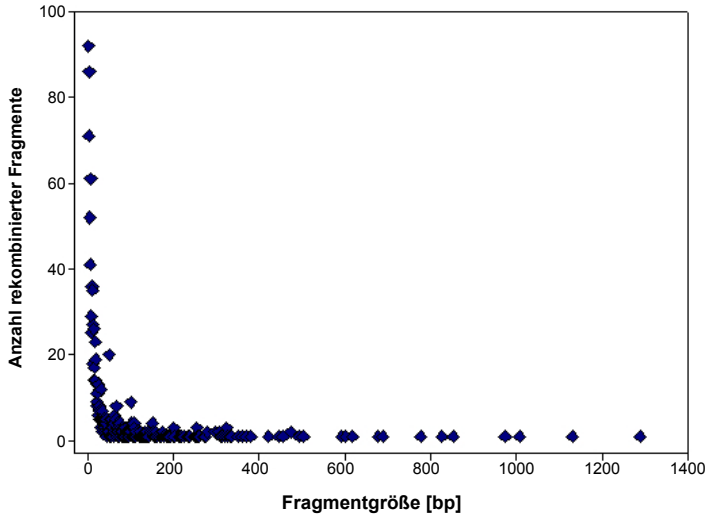

Abbildung 5.12: Größenverteilung der rekombinierten Fragmente in Sa3*int*-Prophagen.

Rekombination entlang des Phagengenoms. Da für die Sa3*int*-Prophagen die meisten Daten vorliegen, werden Untersuchungen zur Rekombination entlang des Phagengenoms an dieser Gruppe untersucht.

Die Anzahl der Rekombinationsereignisse wird gegen die Positionen des Referenzphagen ϕN315 aufgetragen. In Abbildung 5.13 ist gut zu sehen, dass alle Module von Rekombination betroffen sind. Lediglich das Lysogenie-Modul mit der Integrase zeigt weniger Rekombination.

Die Annahme, dass Module als ganzes ausgetauscht werden, wird durch die vorliegenden Daten nicht bestätigt. Es ist außerdem nicht klar, ob Gengrenzen - wie von Clark *et al.* 2001 beschrieben - wirklich bevorzugte Stellen für Rekombination sind, da auch Ereignisse innerhalb von Genen identifiziert wurden.

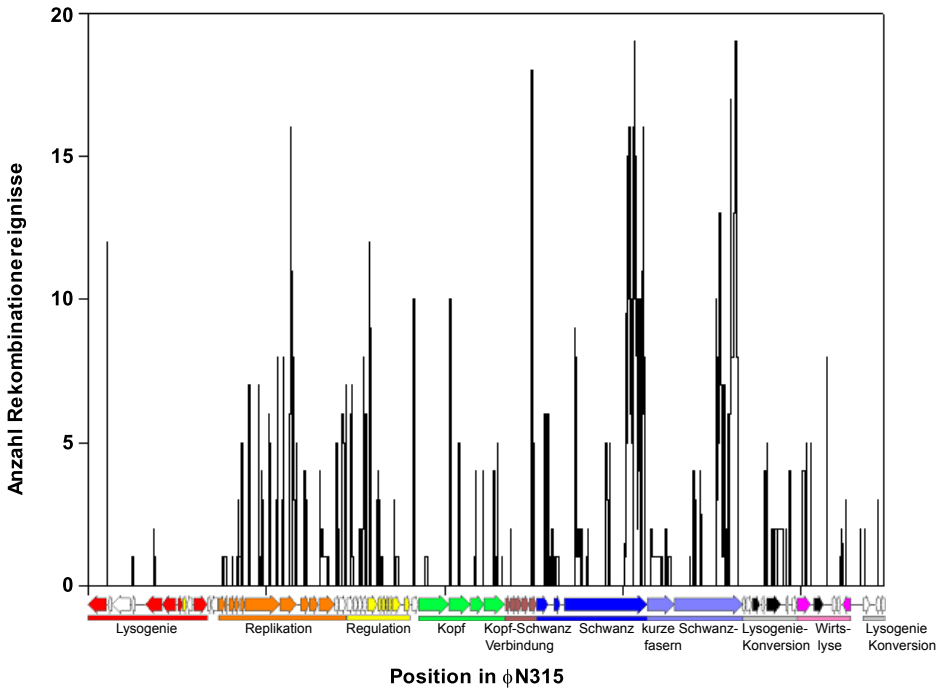

Abbildung 5.13: Rekombination entlang des Prophagen ϕN315. Die Farben der Gene entsprechen dem zugehörigen funktionellen Modul.

Herkunft der rekombinierten Fragmente. Die Sequenzen der rekombinierten Sequenzen werden mittels blastn gegen GenBank geblastet, um eine Aussage über die Herkunft der rekombinierten Fragmente machen zu können. Obwohl Prophagen theoretisch Zugang zu einem globalen Pool anderer Prophagen-Sequenzen haben sollten, hängt die Zugangshäufigkeit stark von der Anzahl verschiedener Barrieren, wie z.B. einem eingeschränktes Wirtsspektrum, ab (Hendrix 1999).
Für die Prophagen des klonalen Komplexes CC5 haben die Untersuchungen gezeigt, dass alle rekombinierten Fragmente aus anderen *S. aureus* Prophagen stammen und es wahrscheinlich einen begrenzten *S. aureus* Prophagen-Pool gibt.

5.4 Mikroevolution des Sequenztyps ST225

5.4.1 Verwendete Isolate

Zur Aufklärung der globalen Populationsstruktur des Sequenztyps ST225 sowie der Rekonstruktion eines Krankenhausausbruchs mit dem Sequenztyp ST225 in Kopenhagen, Dänemark wurden die Genome von 67 Isolaten sequenziert. Das Taxonsampling umfasst Isolate aus vier europäischen Ländern, einem dänischen Krankenhaus, aus verschiedenen Krankenstationen des Krankenhauses sowie eine Serie von Isolaten aus fünf verschiedenen Patienten, für die epidemiologische Daten vorliegen. Tabelle 5.7 enthält eine Übersicht der verwendeten Stämme sowie einige epidemiologische Daten.

Tabelle 5.7: Übersicht der verwendeten ST225 Genome.

Isolat (RKI-ID)	Original-ID	Sequenzier-methode	Anteil Ns [%]		Herkunfts-land	Ort	Kranken-station	Datum der Isolation
01-04209-1		Solexa, GATC	8,13		Deutschland	Oberhausen		20.12.2001
03-02595		Solexa, GATC	3,57		Deutschland	Heidelberg		18.11.2003
04-00194-2		Solexa, GATC	22,12		Deutschland	Heidelberg		20.01.2004
04-02981		454 (Branford), Solexa (GATC)	0,00		Deutschland	Köln		04.10.2004
05-00043		Solexa, GATC	22,10		Deutschland	Freiburg		17.12.2004
05-02010		Solexa, GATC	47,60		Deutschland	Saarbrücken		03.08.2005
06-01100		Solexa, GATC	0,03		Deutschland	Göttingen		23.05.2006
06-01602		454, Göttingen	5,61		Deutschland	Blankenburg		01.08.2006
06-02150		Solexa, GATC	28,61		Deutschland	Rheinbach		16.10.2006
07-00265		454, Göttingen	0,85					25.01.2007
07-00952		Solexa, GATC	0,04		Deutschland	Ingolstadt		30.03.2007
07-01593		Solexa, GATC	0,03		Tschechien	Budweis		13.11.2006
07-01606		Solexa, GATC	0,03		Tschechien	Königgrätz		08.02.2007
07-03027	M124	Solexa, GATC	8,36	Indexpatientin	Dänemark	Kopenhagen	Intensivstation	21.10.2004
07-03028	M145	Solexa, GATC	11,31		Dänemark	Kopenhagen	Ambulanz	30.11.2004
07-03030	M597	Solexa, GATC	0,01		Dänemark	Kopenhagen	Allgemein Medizin (Arztpraxis)	18.08.2006
07-03031	M615	Solexa, GATC	49,88		Dänemark	Kopenhagen	Geriatrie	13.09.2006
07-03032	M704	Solexa, GATC	4,87		Dänemark	Kopenhagen	Geriatrie	14.02.2007
07-03033	M176	Solexa, GATC	17,78		Dänemark	Kopenhagen	Innere Medizin	12.01.2005
07-03034	M126	Solexa, GATC	13,09		Dänemark	Kopenhagen	Intensivstation	27.10.2004
07-03329		Solexa, GATC	1,45		Tschechien	Prag		06.12.2002
07-03330		Solexa, GATC	0,02		Tschechien	Aussig		10.02.2003
07-03336		Solexa, GATC	0,01		Tschechien	?		??.??.2004
07-03462		Solexa, GATC	6,91		Schweiz	Zürich		24.11.2003
08-00392		454, Göttingen	27,17					12.02.2008
08-00463		454, Göttingen	6,44					20.02.2008
08-01881		Solexa, GATC	7,22		Deutschland	Bad Wildungen		28.07.2008
08-02863		454, Göttingen	67,62					??.??.2000
08-02865		Solexa, GATC	3,25		Deutschland	Berlin		22.06.2000
09-00666		Solexa, GATC	4,57		Deutschland	Riesa		25.01.2009
09-00824		454, Göttingen	1,53					??.07.1994
09-00825		454, Göttingen	2,41					??.12.1994
09-00826		454, Göttingen	3,40					??.03.1995
09-02312		Solexa, GATC	2,78		Deutschland	Heide		19.06.2009
09-03403	M639	Solexa, GATC	0,02	Patient 2	Dänemark	Kopenhagen	Geriatrie	06.10.2006
09-03404	D722	Solexa, GATC	2,53	Patient 2	Dänemark	Kopenhagen		22.11.2006
09-03405	D1410	Solexa, GATC	0,33	Patient 2	Dänemark	Kopenhagen	Notaufnahme	24.03.2009
09-03406	D1453	Solexa, GATC	0,01	Patient 2	Dänemark	Kopenhagen		21.04.2009
09-03407	D1466	Solexa, GATC	0,04	Patient 2	Dänemark	Kopenhagen		21.04.2009
09-03408	D1467	Solexa, GATC	0,02	Patient 2	Dänemark	Kopenhagen		21.04.2009
09-03409	D1469	Solexa, GATC	4,02	Patient 2	Dänemark	Kopenhagen		21.04.2009
09-03410	M893	Solexa, GATC	2,93	Patient 3	Dänemark	Kopenhagen	Ambulanz	15.01.2008
09-03411	M968	Solexa, GATC	0,04	Patient 3	Dänemark	Kopenhagen		22.06.2008
09-03412	D1154	Solexa, GATC	0,03	Patient 3	Dänemark	Kopenhagen	Ambulanz	15.08.2008
09-03413	M1104	Solexa, GATC	0,04	Patient 3	Dänemark	Kopenhagen	Ambulanz	19.11.2008
09-03414	D1338	Solexa, GATC	0,06	Patient 3	Dänemark	Kopenhagen	Ambulanz	23.01.2009
09-03415	M1247	Solexa, GATC	0,06	Patient 3	Dänemark	Kopenhagen	Ambulanz	20.04.2009
09-03416	M1321	Solexa, GATC	1,95	Patient 3	Dänemark	Kopenhagen		29.07.2009
10-03045	M1770	HiSeq, GATC	0,02	Patient 3	Dänemark	Kopenhagen	Ambulanz	03.11.2010
10-03048	D1761	HiSeq, GATC	0,08	Patient 3	Dänemark	Kopenhagen		09.12.2009
10-03037	D675	HiSeq, GATC	0,03	Patient 4	Dänemark	Kopenhagen	Geriatrie	19.10.2006
10-03039	M862	HiSeq, GATC	0,02	Patient 4	Dänemark	Kopenhagen		02.11.2007
10-03043	D660	HiSeq, GATC	0,02	Patient 4	Dänemark	Kopenhagen	Geriatrie	05.10.2006
10-03044	D661	HiSeq, GATC	0,03	Patient 4	Dänemark	Kopenhagen	Geriatrie	05.10.2006
10-03052	M633	HiSeq, GATC	0,03	Patient 4	Dänemark	Kopenhagen	Geriatrie	03.10.2006
10-03053	M752	HiSeq, GATC	0,02	Patient 4	Dänemark	Kopenhagen		15.05.2007
10-03056	D673	HiSeq, GATC	0,03	Patient 4	Dänemark	Kopenhagen	Geriatrie	19.10.2006
10-03057	D674	HiSeq, GATC	0,03	Patient 4	Dänemark	Kopenhagen	Geriatrie	19.10.2006
10-03038	D1743	HiSeq, GATC	0,02	Patient 5	Dänemark	Kopenhagen	Orthopädie	25.11.2009
10-03040	M885	HiSeq, GATC	0,02	Patient 5	Dänemark	Kopenhagen	Innere	29.12.2007
10-03041	M967	HiSeq, GATC	0,02	Patient 5	Dänemark	Kopenhagen	Innere	19.06.2008
10-03042	M1429	HiSeq, GATC	0,02	Patient 5	Dänemark	Kopenhagen	Orthopädie	11.11.2009
10-03046	D1157	HiSeq, GATC	0,03	Patient 5	Dänemark	Kopenhagen		11.08.2008
10-03054	D1155	HiSeq, GATC	0,02	Patient 5	Dänemark	Kopenhagen		11.08.2008
10-03055	D1156	HiSeq, GATC	0,03	Patient 5	Dänemark	Kopenhagen		11.08.2008
10-03049	H2	HiSeq, GATC	0,02	Patient 6	Dänemark	Kopenhagen	nicht hospitalisiert	10.12.2004
10-03050	H363	HiSeq, GATC	0,02	Patient 6	Dänemark	Kopenhagen	Neurologie	01.12.2006
10-03051	H6	HiSeq, GATC	0,02	Patient 6	Dänemark	Kopenhagen	Neurologie	27.04.2005

5.4.2 Sequenzierung und Readmapping

Die Genome der verwendeten Isolate wurden mittels Solexa (39), Solexa HiSeq (20) und 454 (8) sequenziert. Die Isolate 05-02010, 07-03031 und 08-02863 haben nach dem Mapping auf die Referenzsequenz 04-02981 und der anschließenden Qualitätsüberprüfung der einzelnen Basen jeweils einen Gesamtanteil von > 47 % nicht aufgelöster Nukleotide („N“s). Aufgrund dieser schlechten Werte werden diese drei Isolate von den folgenden Analysen ausgeschlossen.

5.4.3 Globale Populationsstruktur

Die Genom-basierte Rekonstruktion der globalen Populationsstruktur des Sequenztyps ST225 beruht auf 36 internationalen Isolaten. Als Außengruppe wird der Stamm JH1 (Sequenztyp ST105) verwendet.
Nach dem Ausschluss von SNPs in Wiederholungssequenzen und mobilen genetischen Elementen, nicht aufgelösten Nukleotiden („N"s), Alignmentlücken und Positionen mit schlechter Sequenzier- und Mappingqualität besitzt das Alignment aus variablen Positionen eine Länge von 326 bp (davon 46 SNPs spezifisch für die Außengruppe). Für die Daten ist das TVM-Modell das am besten für die Rekonstruktion des „*Maximum Likelihood*"-Stammbaums geeignete Substitutionsmodell. Abbildung 5.14 zeigt die Phylogenie des Sequenztyps ST225.
Die Isolate aus Deutschland, Tschechien und den USA clustern je nach geographischer Herkunft zusammen. Die dänischen Isolate dagegen bilden zwei nicht nicht verwandte Cluster (Kapitel 5.4.4). Die US-amerikanischen Stämmen nehmen eine basale Stellung ein und bestätigen noch einmal den nordamerikanischen Ursprung des Sequenztyps ST225 (Nübel *et al.* 2010, Kapitel 5.2.3). Innerhalb der europäischen Stämme sind die Isolate aus Deutschland basal gegenüber den Isolaten aus Tschechien, Dänemark und der Schweiz. Auch diese Beobachtung ist kongruent mit dem Ergebnis einer Analyse zur zeitlichen und räumlichen Ausbreitung des Sequenztyps ST225. Dabei konnten Nübel *et al.* (2010) zeigen, dass sich der Sequenztyp ST225 von Deutschland aus in Europa verbreitet hat.

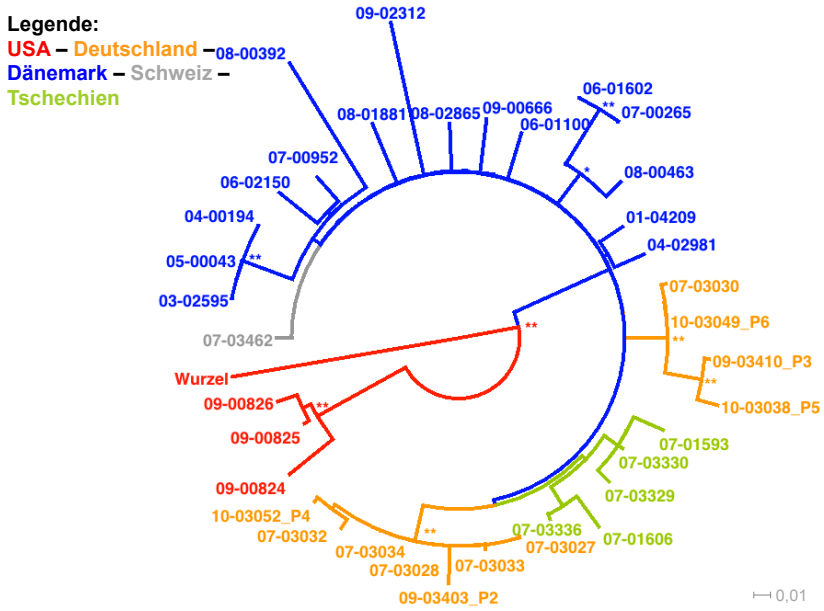

Abbildung 5.14: Globale Populationsstruktur des Sequenztyps ST225 basierend auf 326 SNPs. ** approximierter Boostrap-Wert = 100 * approximierter Boostrap-Wert = 95 - 99. Der Maßstab zeigt die evolutionäre Distanz in Substitutionen pro Position.

5.4.4 Rekonstruktion eines Ausbruchs im Krankenhaus

Zur Rekonstruktion eines MRSA-Ausbruchs in einem Krankenhaus in Kopenhagen, Dänemark werden 13 Stämme verwendet. Abbildung 5.15 zeigt die phylogenetischen Zusammenhänge der dänischen Krankenhaus-Stämme.

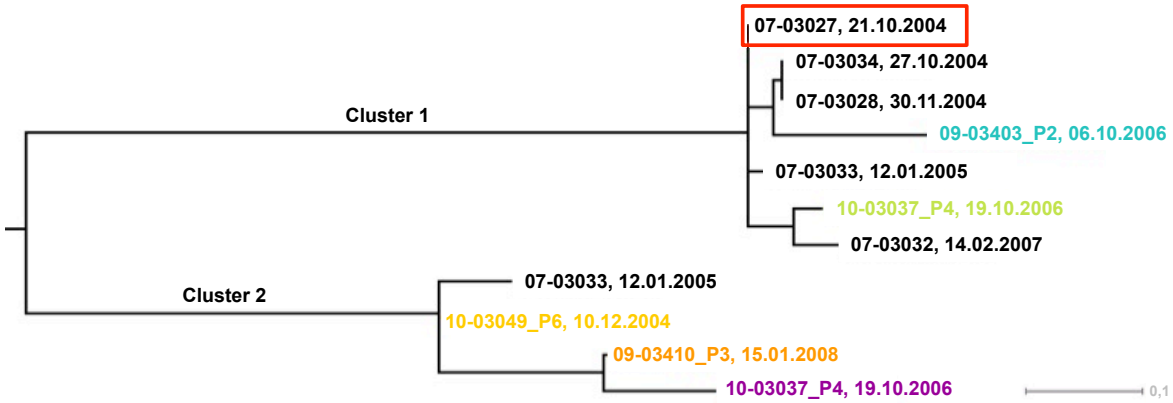

Abbildung 5.15: Rekonstruktion eines Ausbruchs mit dem Sequenztyp ST255 in einem dänischen Krankenhaus. „*Maximum Likelihood*"-Baum von 13 MRSA-Stämmen basierend auf 80 SNPs. Der Maßstab zeigt die evolutionäre Distanz in Substitutionen pro Position.

Es sind deutlich zwei Cluster zu unterscheiden. Cluster 1 enthält das Isolat 07-03027. In einer früheren Studie konnte die Index-Patientin identifiziert werden, die 2004 diesen Klon des Sequenztyps ST225 aus einem Krankenhaus in Frankfurt, Deutschland in das dänische Krankenhaus eingeschleppt hat (Nübel *et al.* 2010).
Das zweite Cluster repräsentiert einen zweiten Ausbruch, der unabhängig von dem oben beschriebenen Index-Fall ist. Die ähnliche Länge der Äste deutet darauf hin, dass ein weiterer Stamm des Sequenztyps ST225 etwa zur gleichen Zeit in das Krankenhaus gekommen ist, der aber mit den verwendeten Typisierungsmethoden nicht unterschieden werden konnte.

5.4.5 Evolution der Isolate der Patienten-Reihe

Um einen Einblick in die Evolution von MRSA in einzelnen Patienten zu bekommen, wurden 34 MRSA-Stämme aus insgesamt fünf Patienten sequenziert, die über mehrere Jahre isoliert wurden. Für diese Patienten sind Daten über Zeitpunkt und Ort der Isolation sowie der Antibiotika-Therapien vorhanden.
Abbildung 5.16 zeigt die Phylogenie der Patientenstämme, die auf einem Alignment aus 161 variablen Positionen beruht (10 SNPs spezifisch für die Außengruppe 04-02981). Neben den Patienten-Isolaten sind auch die fünf zusätzlichen Stämme aus dem Krankenhaus im Stammbaum enthalten. Tabelle A.2 im Anhang enthält eine Übersicht der SNPs in den Patientenstämmen. Einige Patienten-spezifische SNPs wurden per PCR verifiziert. Die verwendeten Primer sind in Tabelle A.3 im Anhang aufgelistet.

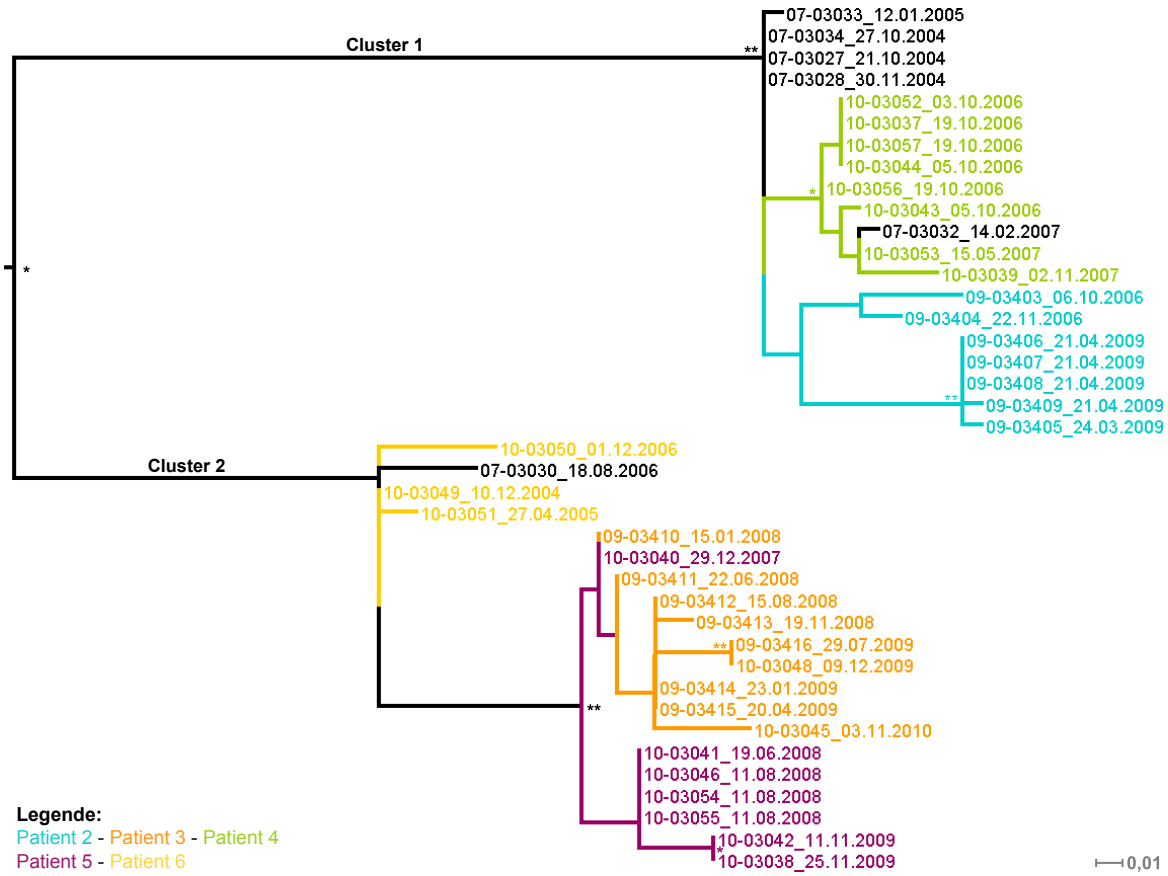

Abbildung 5.16: „*Maximum Likelihood*"-Baum der Patienten-Isolate basierend auf 161 SNPs. ** LR-ELWs = 100 * LR-ELWs = 95 - 99. Der Maßstab zeigt die evolutionäre Distanz in Substitutionen pro Position.

Die Isolate der fünf Patienten clustern - wie schon in Abbildung 5.15 zu sehen - in zwei Gruppen zusammen, die zueinander eine starke Diversität aufweisen (56 SNPs diversifizieren die beiden Cluster).

Cluster 1 umfasst die Isolate der Patienten 2 und 4 sowie die Stämme 07-03027, 07-03028, 07-03033 und 07-03034. Diese vier Stämme nehmen die basale Stellung im Cluster ein. Das Isolat 07-03027 wurde bereits von Nübel *et al.* 2010 als Index-Isolat identifiziert. Die epidemiologischen Daten zeigen, dass die zugehörigen Patienten zeitweise auf der gleichen Station lagen. Die Stämme des Patienten 2 bilden ein monophyletisches Cluster. Das Cluster von Patient 4 ist dagegen paraphyletisch, da es auch das Isolat 07-03032 enthält. Auch hier gibt es einen epidemiologischen Zusammenhang, da Patient 4 und der Patient mit dem Isolat 07-03032 beide zur gleichen Zeit in der Geriatrie behandelt wurden.

Im zweiten Cluster sind keine monophyletischen Gruppen enthalten. Die basale Klade enthält die Isolate des Patienten 6, die paraphyletisch zu dem Isolat 07-03030 sind. Die Patienten lagen gleichzeitig in der selben Krankenhausstation (persönliche Mitteilung Mette Bartels; nähere Informationen liegen zur Zeit nicht vor). Die Patientenstämme 3

und 5 bilden zwei Schwestergruppen zueinander und sind paraphyletisch. Das Isolat 10-03040 im Cluster des Patienten 3 gehört eigentlich zu Patient 5. Zwischen diesen beiden Patienten gibt es keinen offensichtlichen epidemiologischen Zusammenhang; allerdings leben beide in der gleichen Gegend in Kopenhagen nur wenige Straßen voneinander entfernt.

5.4.6 Evolution innerhalb von Patienten

Werden die Stammbäume der Patienten einzeln betrachtet, erhöht sich im Laufe der Zeit die genetische Distanz zwischen den Isolaten (Abbildung 5.17). Da - mit Ausnahme von Patient 4 - alle Patienten eine Zeit lang und zum Teil wiederholend mit Antibiotika behandelt wurden, könnte ein Zusammenhang zwischen der Diversifizierung und der Behandlung angenommen werden. Allerdings ist die Distanz der Isolate bzw. die Anzahl der angehäuften SNPs zur Wurzel sowohl zwischen den Patienten mit Behandlung als auch im Vergleich zu dem Patienten ohne Behandlung ähnlich und eine Korrelation zwischen Diversifizierung und Therapie ist nicht gegeben.
Insgesamt sind 58 % der Mutationen in Cluster 1 und Cluster 2 nicht-synonym; der Anteil synonymer bzw. intergenischer SNPs liegt bei 22 bzw. 20 %. Einige der nicht-synonymen Mutationen sind in Genen zu finden, die mit der Virulenz von *S. aureus* im Zusammenhang stehen. Allerdings sind diese in fast allen Fällen spezifisch für ein einzelnes Isolat. Das Isolat 10 -03039 des Patienten 4 ist das einzige Isolat, das eine nicht-synonyme Mutation in einem Gen (*rpoB*) besitzt, welches mit der Ausbildung einer Anitbiotika-Resistenz (Rifampicin) assoziiert ist. Allerdings ist für die vorhandene Mutation kein Zusammenhang zur Resistenz beschrieben (Aubry-Damon, Soussy & Courvalin 1998) und der Patient wurde auch nicht mit Antibiotika behandelt. Ein ausgeprägter positiver Selektionsdruck scheint nicht auf die betroffenen Gene zu wirken.
Einige Stämme wurden zeitgleich, aber an verschiedenen Körperstellen isoliert (Patient 2, 21.04.2009; Patient 4, 19.10.2006; Patient 5, 11.08.2008). Die Stammbäume der Patientenisolate zeigen keine bis kaum Diversität innerhalb dieser Stämme (Abbildung 5.17).

5.4.7 Statistiken

Verteilung von SNPs. Die SNPs der untersuchten Isolate des Sequenztyps ST225 sind gleichmäßig über das Kerngenom verteilt. Obwohl SNPs in MGEs für die phylogenetischen Analysen ausgeschlossen wurden, soll auf diese der Vollständigkeit halber kurz eingegangen werden.
Wie auch schon für den klonalen Komplex CC5 beobachtet, treten Sequenzunterschiede vor allem in Prophagen auf, wogegen die übrigen mobilen genetischen Elementen kaum von Mutationen oder Indels betroffen sind. Populationsweit und zwischen den

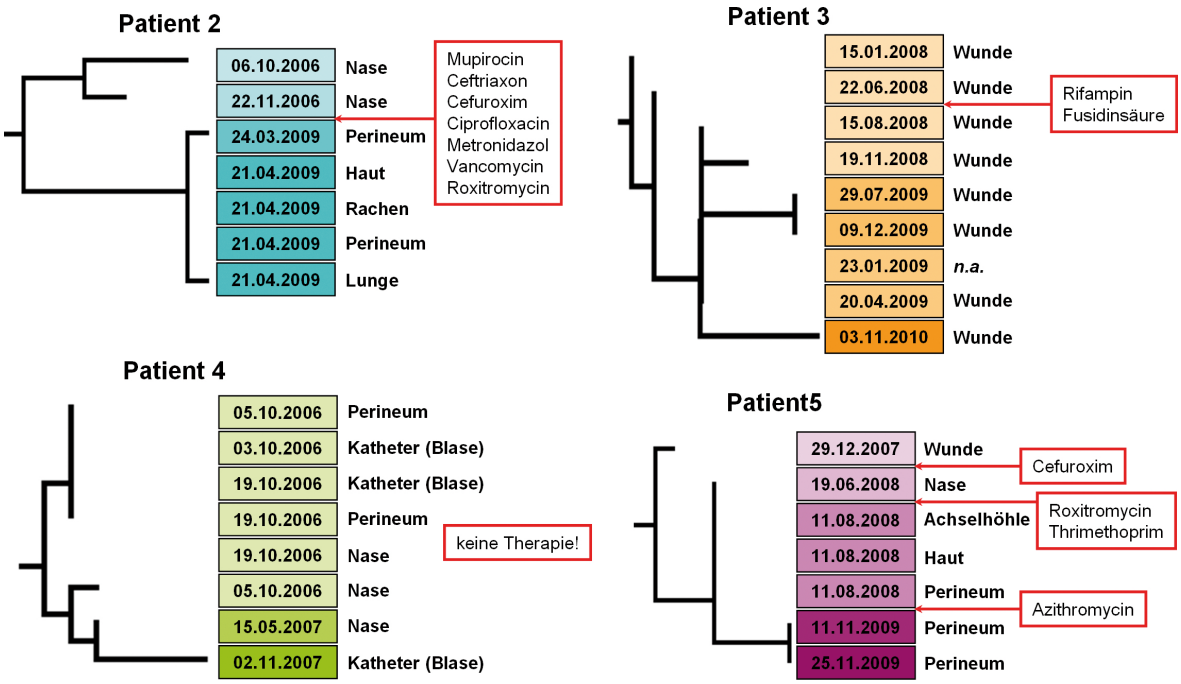

Abbildung 5.17: Evolution von Isolaten in einzelnen Patienten. Angegeben sind das Isolationsdatum, die Entnahmestellen der Isolate sowie die eingesetzten Therapeutika. Der Farbcode symbolisiert die Distanz zwischen den Isolaten.

verschiedenen Patienten ist die Häufung von Unterschieden in den Prophagen besonders ausgeprägt; innerhalb einzelner Patienten zeigen die Prophagen dagegen nur wenig Diversität.

Sequenzvariationen. Phylogenetische Analysen sind nur sinnvoll, wenn genug Information, d.h. eine ausreichend große Diversität, im Datensatz vorhanden ist. Um dies für die verwendeten Datensätze zu verifizieren, wird die Distanz zwischen einem Isolat und der Wurzel im Stammbaum gegen den Beprobungszeitpunkt aufgetragen (Abbildung 5.18).

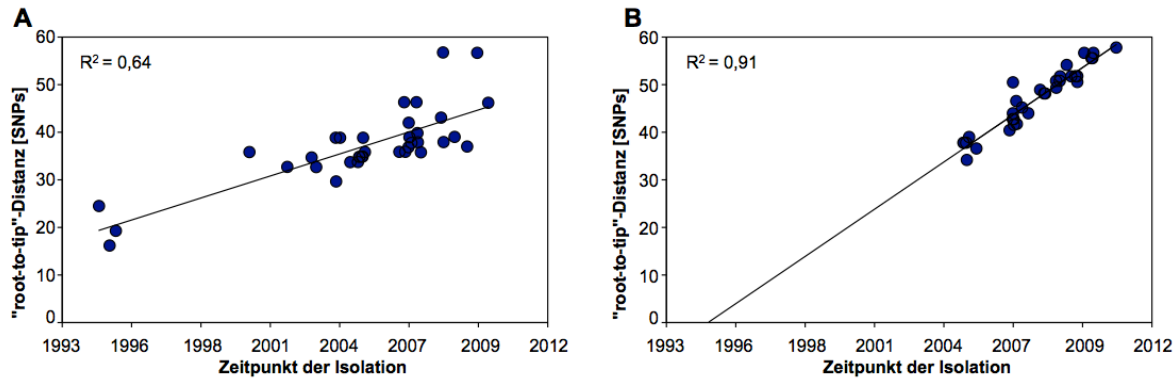

Abbildung 5.18: Zunahme der DNA Sequenzvariation über den Beprobungszeitraum. „Root-to-tip“ Distanz für (A) die Population und (B) die Patienten-Isolate.

Sowohl der Populationsweite- als auch der Patienten-Datensatz zeigen eine positive Korrelation der Divergenz mit dem Beprobungszeitpunkt. Somit ist in beiden Datensätzen ein starkes zeitliches Signal enthalten und eine signifikante Anhäufung von Sequenzunterschieden zu erkennen. Besonders innerhalb der Patienten-Isolate hat die Divergenz in sehr kurzer Zeit stark zugenommen.

Einfluss von Selektion: dN/dS. dN/dS wurde sowohl für den populationsweiten als auch für den Patienten-Datensatz berechnet. Im Mittel ergibt sich ein dN/dS-Wert von 0,6. Genau wie für den klonalen Komplex CC5 ist der Einfluss reinigender Selektion damit nur mäßig vorhanden.

Anhäufung von AT-Mutationen. Der Anteil AT-anhäufender Mutationen (GC $\rightarrow$ AT) wurde sowohl populationsweit als auch für die Patienten berechnet. Für die Patienten-Phylogenie wurden zusätzlich auch die Werte für die weiter innenliegenden (älteren) Äste ermittelt.
Populationsweit liegt der Wert +AT/+GC bei 4 und für den Patienten-Datensatz bei 1,8. Abbildung 5.19 zeigt für den Patienten-Datensatz sowohl den Anteil an AT-Mutationen als auch dN/dS gegen dS, wobei dS als Maß für das Alter der Isolate dient.
Deutlich zu sehen ist, dass die Anzahl an AT-Mutationen abnimmt, je älter ein Ast im phylogenetischen Baum ist. Diese Beobachtung korreliert mit den dN/dS-Werten. Die Anzahl nicht-synonymer SNPs ist zwischen nah verwandten Isolaten noch hoch, nimmt jedoch mit steigender Distanz ab und dN/dS pendelt sich auf einen Wert um 0,6 ein.

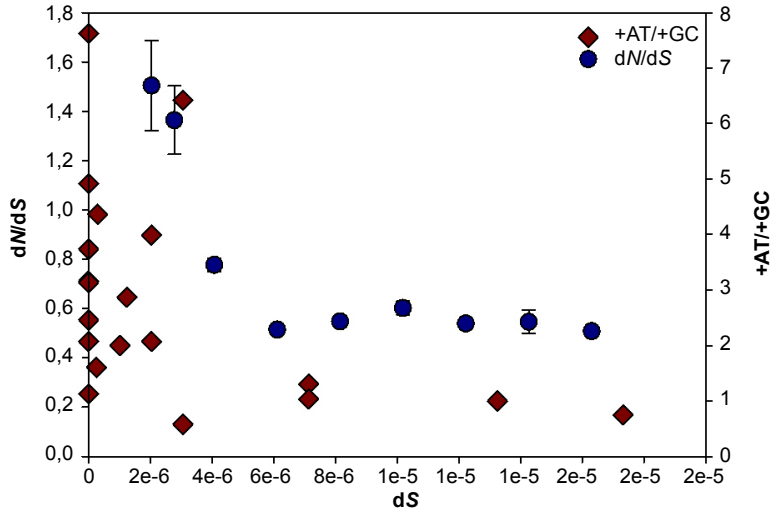

Abbildung 5.19: Anzahl AT-anhäufender Mutationen bzw. dN/dS gegen dS im Patienten-Datensatz.

Homoplasie. Homoplastische SNPs wurden weder im Populations- noch im Patienten-Datensatz detektiert. Der Homoplasie-Index ist somit 0,00.

5.4.8 Raten und Daten

Evolutionsraten. Die Berechnung der Evolutionsraten erfolgte mittels BEAST (Drummond & Rambaut 2007) und wurde für einen klonalen Komplex, verschiedene Sequenztypen und fünf verschiedene Patienten eines Sequenztyps durchgeführt (Abbildung 5.20). Die Abbildung 5.20 zeigt, dass es keinen signifikanten Unterschied in den Evolutionsraten in den verschiedenen Populationsebenen gibt.

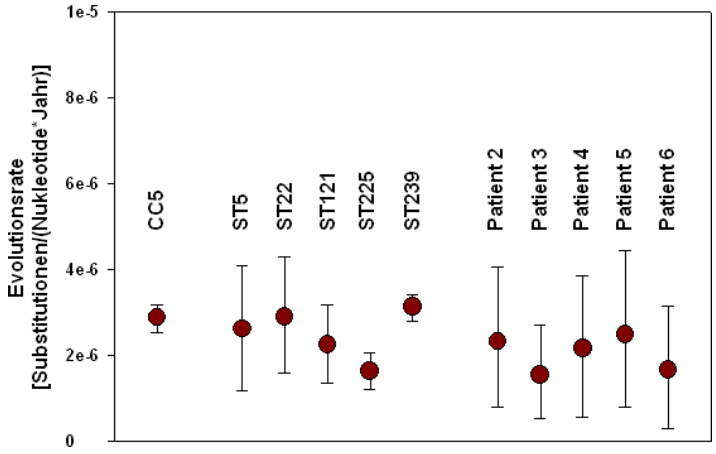

Abbildung 5.20: Evolutionsraten verschiedener Sequenztypen bzw. Diversitätsleveln. Die Daten für ST239 sind Gray *et al.* (2011) entnommen; die Daten für CC5, ST5, ST22 und ST121 sind persönliche Mitteilungen von U. Nübel.

Datierung des Besiedlungszeitpunkts. Aus der von BEAST (Drummond & Rambaut 2007) berechneten Zeit seit dem letzten gemeinsamen Vorfahren (engl. „*time to most recent common ancestor*“, TMRCA) und den Zeitpunkten der Beprobung lässt sich für die fünf Patienten der Zeitpunkt der Besiedlung/Infizierung berechnen. Tabelle 5.8 zeigt die Ergebnisse dieser Berechnung. Weder die Mittelwerte noch die Konfidenzintervale der wahrscheinlichen Entstehungszeitpunkte zeigen Konflikte mit dem jeweils ersten Isolationsdatum jedes Patienten.

Tabelle 5.8: Datierung von Infektionen.

Patient	Anzahl Isolate	Erstes Isolat	Entstehung	
			Mittelwert	95 % Konfidenzintervall
Patient 2	7	2006	2006	2004 - 2006
Patient 3	9	2008	2005	2005 - 2007
Patient 4	8	2006	2006	2005 - 2006
Patient 5	7	2007	2007	2006 - 2007
Patient 6	3	2004	2003	2001 - 2004

5.5 Skripte

Für diese Arbeit wurden einige Perl- und Python-Scripte entwickelt bzw. schon vorhandene angepasst. Tabelle 5.9 zeigt eine Übersicht.

Tabelle 5.9: Übersicht der entwickelten und angepassten Scripte.

Name	Sprache	Anwendung	Entwickler
Consensus.pl	Perl	Ändert unspezifische Nukleobasen in die Base einer Referenzsequenz um	„*in-house*“
ClonalFrameParser.pl	Perl	Ausgabe rekombinierter Fragmente mit Start- und Endposition bei bestimmter „*posterior probability*“ (A-Posteriori-Wahrscheinlichkeit)	„*in-house*“
FastQ-Converter (diverse)	Perl	Konvertieren von *.fasta und *.qual in *.fastq	SEQanswers Forum[3], Anpassungen „*in-house*“
QualityFilter.py	Python	Nukleotide mit schlechten Qualitätswerten werden durch ein „N“ ersetzt	„*in-house*“
RegionRemover.pl	Perl	Entfernen von vorgegebenen Sequenzabschnitten in einer *.fasta Datei	„*in-house*“
Repair_MatePair.py	Python	Verarbeitung von fehlerhaften „*paired-end*“-fastq-Dateien	SEQanswers Forum[4]

[3]Entwickler: kmcarr, URL: `http://seqanswers.com/forums/showpost.php?p=7248&postcount=17`; Stand: 29.11.2011

[4]Entwickler: maubp, URL: `http://seqanswers.com/forums/showpost.php?p=22595&postcount=6`; Stand: 29.11.2011

6 Diskussion

6.1 Mikroevolution des klonalen Komplexes CC5

6.1.1 Phylogenie des klonalen Komplexes CC5

Durch den Einsatz von Hochdurchsatz-Sequenziertechnologien und der sich daraus ergebenden Möglichkeit, schnell und kostengünstig ganze Genome zu sequenzieren, konnte die globale Phylogenie des klonalen Komplexes CC5 mit der bisher größten Auflösung rekonstruiert werden.
Statt den von Nübel *et al.* (2008) zur Untersuchung des Sequenztyps ST5 verwendeten 1,5 % des Genoms (46 kb) wird für die vorliegende Analyse, wie bereits erwähnt, das ganze Genom mit ca. 2.800 kb genutzt. Die phylogenetischen Abstammungslinien werden dadurch bestätigt, allerdings sind Informationen über Zweiglängen und Knotenpositionen sehr viel präziser und verwandtschaftliche Beziehungen werden besser aufgelöst.
Die Aufnahme einiger US-amerikanischer Isolate des Sequenztyps ST5 zeigt eine nahe Verwandtschaft zu den US-amerikanischen Isolaten JH1 und JH9 (Sequenztyp ST105, Mwangi *et al.* 2007) sowie dem deutschen Isolat 04-02981 (Sequenztyp ST225), was die Theorie eines nordamerikanischen Ursprungs des Sequenztyps ST225 unterstützt (Nübel *et al.* 2010).

6.1.2 Geringer Einfluss von Selektion auf den klonalen Komplex CC5

Um den Einfluss von Selektion auf nah verwandte Isolate richtig beurteilen zu können, sollten die Ergebnisse einer Studie von Rocha *et al.* (2006) berücksichtigt werden.
In dieser wurde gezeigt, dass $\mathrm{d}N/\mathrm{d}S$ sinkt, je mehr Zeit seit der Diversifizierung von zwei Isolaten vergangen ist. Der Grund dafür ist, dass schädliche Mutationen im Laufe der Zeit durch reinigende Selektion aus dem Genom entfernt werden. Es kommt zu einer Abnahme nicht-synonymer Mutationen in den Genomen.
Betrachtet man $\mathrm{d}N/\mathrm{d}S$ zwischen dem klonalen Komplex CC5 und anderen klonalen Komplexen (Abbildung 5.6B) kann der von Rocha *et al.* (2006) beschriebene starke Effekt reinigender Selektion und die damit verbundene Anhäufung synonymer SNPs beobachtet werden.
Für Protein-kodierende Gene innerhalb des klonalen Komplexes CC5 wurde ein $\mathrm{d}N/\mathrm{d}S$-Wert von 0,51 berechnet. Der Anteil nicht-synonymer Mutationen ist noch recht hoch und es ist nur ein mäßiger Effekt von reinigender Selektion zu beobachten. Seit der Entstehung des klonalen Komplexes CC5 ist also nicht genug Zeit vergangen, um alle schädlichen Mutationen zu entfernen.

6.1.3 Kaum Homoplasien im klonalen Komplex CC5

Der Anteil homoplastischer SNPs im Datensatz des klonalen Komplexes CC5 ist extrem gering: Von den gefundenen 2.971 SNPs sind lediglich sechs homoplastisch (0,2 %; Homoplasie-Index = 0,0041). Die beschriebene Phylogenie ist damit annähernd eindeutig.
Homoplasien können entweder durch Rekombination oder durch konvergente Evolution entstehen. Da Rekombination im Kerngenom von *S. aureus* sehr gering ist (Spratt, Hanage & Feil 2001) und keine Bereiche mit angehäuften SNPs gefunden wurden (die durch Rekombination entstehen würden), sind die vorhandenen Homoplasien eher mit Konvergenz zu erklären. In diesem Fall würde ein positiver Selektionsdruck auf diese Positionen wirken. Allerdings ist für die betroffenen Gene weder ein Zusammenhang in der Ausbildung von Antibiotika-Resistenzen noch mit Interaktionen zwischen Pathogen und Wirt beschrieben.
50 % der homoplastischen SNPs im Datensatz liegen in intergenischen Regionen, obwohl diese Regionen insgesamt lediglich 16 % des *S. aureus* Genoms ausmachen (Rogozin *et al.* 2002). Gemeinhin wird angenommen, dass Mutationen in intergenischer DNA selektiv neutral sind, obwohl phänotypische Auswirkungen vorstellbar sind, da sie z.B. Promotoren, Terminatoren und andere Transkriptionssignale enthalten. SNPs in diesen Bereichen können starke Effekte auf Genexpression, Physiologie und Virulenz von *S. aureus* haben (Villaruz *et al.* 2009).
Zusätzlich werden in Bereichen zwischen Genen nicht-kodierende RNAs (ncRNAs) transkribiert, die an vielen zellulären Prozessen beteiligt sind. So regulieren sie u.a. auch die Translation. Allerdings liegt keiner der im klonalen Komplex CC5 gefundenen homoplastischen SNPs innerhalb für *S. aureus* bekannter ncRNAs (Geissmann *et al.* 2009, Marchais *et al.* 2009, Pichon & Felden 2005, Abu-Qatouseh *et al.* 2010, Bohn *et al.* 2010).

6.1.4 Vergleichende Genomik: CC5 ist wenig variabel - Ausnahme: Prophagen

Harris *et al.* publizierten 2010 den ersten großen Datensatz aus 63 *S. aureus*-Genomen, um die Diversität eines Sequenztyps aufzuklären. Basierend auf dem gleichen Datensatz veröffentlichten Castillo-Ramírez *et al.* (2011) eine Arbeit, in der sie den Einfluss von Rekombination auf das Kern- und das akzessorische Genom untersuchten. Diese Publikationen beruhten im Wesentlichen auf der Analyse von SNPs. In der vorliegenden Arbeit wurde zum ersten Mal die genetische Ausstattung einer größeren Anzahl Isolate eines klonalen Komplexes miteinander verglichen, die nicht nur auf SNPs sondern auch auf der *de novo* Assemblierung der mobilen genetischen Elemente beruht.
Das Kerngenom weist die erwartete geringe Variabilität auf, die bereits beschrieben

wurde (Lindsay & Holden 2004). Interessanterweise zeigen aber einige der mobilen genetischen Elemente, die in *S. aureus* das akzessorische Genom bilden, ebenfalls wenig Diversität. Die Pathogenitätsinseln SaPI2 und SaPI3 weisen jeweils nur geringe Unterschiede auf und kommen in allen untersuchten Isolaten vor; SaPI1 ist spezifisch für die ostasiatische Linie, aber in diesen Isolaten ebenfalls wenig variabel. Die Typisierung der SCC*mec*-Elemente hat gezeigt, dass im klonalen Komplex CC5 verschiedene Typen vorkommen, was bereits mehrfach beschrieben wurde (Nübel *et al.* 2008, Monecke *et al.* 2011).
Die Analysen der in das Genom integrierten Prophagen enthüllten dagegen eine große Diversität, weswegen sich ein Teil dieser Arbeit mit der näheren Betrachtung von Prophagen beschäftigt.

6.2 Prophagen im klonalen Komplex CC5

Vor 40 Jahren wurden die ersten wissenschaftlichen Arbeiten veröffentlicht, die sich mit der Frage beschäftigten, wie sich die enorme genetische Diversität von Phagen entwickelt (Westmoreland, Szybalski & Ris 1969, Simon, Davis & Davidson *et al.* 1971, Short & Suttle 1999). Obwohl in den letzten Jahren die Genome vieler Prophagen sequenziert wurden, sind viele Fragen zur Evolution von Phagen nach wie vor unbeantwortet.
Im klonalen Komplex CC5 sind Prophagen das am häufigsten vorkommende und diverseste mobile genetische Element. Mit 58 Prophagen-Sequenzen liegt in dieser Arbeit der *bis dato* größte untersuchte Prophagen-Datensatz für einen klonalen Komplex von *S. aureus* vor. Zudem sind die Sequenzen der Wirtsgenome vorhanden und damit ein zeitlicher Zusammenhang zwischen bakteriellem Wirt und Prophagen. So ist es möglich, die Diversität und molekulare Mechanismen der Phagenevolution mit größtmöglicher Auflösung zu untersuchen.

6.2.1 Klassifizierung der Prophagen

Die Klassifizierung von Prophagen ist aufgrund der hohen Diversität und den durch Rekombination entstehenden Mosaikstrukturen des Genoms sehr schwierig (Canchaya *et al.* 2003, Kwan *et al.* 2005). Goerke *et al.* (2009) etablierten ein Klassifizierungsschema, welches auf der Sequenz des Integrase-Gens beruht und das für die in dieser Arbeit verwendeten Prophagen angewendet wurde.
Goerke *et al.* (2009) konnten sieben vorwiegend in der *S. aureus* Prophagen-Population vorkommende Integrase-Gruppen identifizieren, die sowohl von Goerke *et al.* als auch im vorliegenden Datensatz alle im klonalen Komplex CC5 gefunden wurden. Der prozentuale Anteil der verschiedenen Gruppen ist mit wenigen Ausnahmen vergleichbar. Sa3*int* ist für beide Datensätze der am häufigsten vorkommende Integrase-Typ; der

zweithäufigste Typ ist bei Goerke *et al.* Sa7*int* wogegen er im Arbeits-Datensatz Sa1*int* ist. Dieser Unterschied kann mit der geographischen Herkunft der bakteriellen Isolate erklärt werden: Goerke *et al.* untersuchten lediglich Stämme aus Deutschland (plus einigen aus der Schweiz). Sa1*int*-Prophagen entstammen in dieser Arbeit vor allem den US-amerikanischen Isolaten, die einen großen Teil des Taxonsamplings ausmachen.
Da für diese Arbeit komplette Genomsequenzen der Prophagen vorliegen, wurden zwei weitere Ansätze zur Klassifizierung angewendet: Der erste beruht auf der Nukleotidsequenz und der zweite auf dem Phagen Proteom. Beide Methoden zeigen ein nahezu identisches Ergebnis und die Cluster entsprechen - mit wenigen Ausnahmen - den verschiedenen Integrase-Gruppen. Es ist daher davon auszugehen, dass zwischen den *int*-Gruppen wenig Austausch von Sequenzabschnitten stattfindet. Vor allem die Integrase-Typen, die mit einer ausreichend großen Anzahl vertreten sind, geben die Verwandtschaft der Prophagen gut wieder. Eine Ausnahme sind die Prophagen Sa1*int*_ED98 und Sa3*int*_ED98, die nicht in ihrer entsprechenden Sa*int*-Gruppe enthalten sind. Dies spiegelt sehr gut die unterschiedliche Evolution des Wirts dieses Prophagen wieder: ED98 wurde aus einem Hähnchen isoliert und unterlag damit ganz anderen Evolutionskräften als die humanen Isolate. Lowder *et al.* (2009) haben gezeigt, dass der Klon mit mobilen genetischen Elementen ausgestattet ist, die nur in Geflügel-Isolaten vorkommen.
Die drei verwendeten Schemata sind je nach Fragestellung unterschiedlich gut geeignet, um Prophagen zu klassifizieren. Die Verwendung des *int*-Gens anhand der Sequenzierung eines PCR-Produkts ist einfacher, schneller und kostengünstiger als die Sequenzierung eines kompletten Genoms. Die große Diversität innerhalb einer *int*-Gruppe kann allerdings nicht aufgelöst werden, wenn nur das *int*-Gen sequenziert wird. Auch reicht die diskriminatorische Fähigkeit nicht aus, um Prophagen unterscheiden zu können, deren Wirte eine völlig andere evolutionäre Geschichte durchlaufen haben (z. B. *S. aureus* aus Geflügel). Liegen ganze Genomsequenzen der Phagen vor, sollten diese vollständig verwendet werden, um die komplette Diversität abbilden zu können. Einfach und schnell ist die *in silico*-Analyse von Nukleotidsequenzen, wenn genug Sequenzen und damit viele Zwischenstufen der Entwicklung der Diversität vorliegen. Ist dies nicht der Fall, ist eine Alignierung der DNA-Sequenzen aufgrund ihrer geringen Ähnlichkeit schwierig durch zu führen. In so einem Fall kann die Klassifizierung anhand des Proteoms der Prophagen nützlich sein (Rohwer & Edwards 2002, Kwan *et al.* 2005). Auf diese Weise können verwandtschaftliche Beziehungen von Genen berücksichtigt werden, deren Aminosäuresequenzen sich ähneln. Nachteilig an dieser Methode ist der recht große Aufwand der *in silico*-Analysen.

6.2.2 Ausgeprägte Mosaikstrukturen in Prophagen des klonalen Komplexes CC5

Die Prophagen des klonalen Komplexes CC5 zeichnen sich sowohl durch eine große Anzahl SNPs als auch durch Abschnitte mit starken Sequenzunterschieden aus. Mit dem Programm ClonalFrame (Didelot & Falush 2007) wurden in den hier vorliegenden Prophagen-Genomen Rekombinationsereignisse und deren Positionen im Genom detektiert.
Die Analysen haben gezeigt, dass Rekombination nahezu gleichmäßig über das Prophagen-Genom verteilt auftritt und alle funktionellen Module betroffen sind (Abbildung 5.13). Die Theorie, dass funktionelle Module als ganzes ausgetauscht werden (vgl. Mykobakteriophagen: Pedulla *et al.* 2003), kann durch die vorliegenden Daten nicht bestätigt werden. Unklar ist, ob Gengrenzen - wie von Clark *et al.* 2001 beschrieben - wirklich bevorzugte Stellen für Rekombination sind, da auch zahlreiche Ereignisse innerhalb von Genen identifiziert wurden.
Das hohe Maß an Sequenzunterschieden sowie die Ergebnisse der Rekombinationsanalysen sprechen für einen mosaikartigen Aufbau des Prophagen-Genoms. Zum ersten Mal kann für einen definierten Datensatz gezeigt werden, dass bereits innerhalb eines klonalen Komplexes von *S. aureus* die Mosaikstrukturen der Prophagen sehr stark ausgeprägt sind. Frühere Analysen haben einen mosaikartigen Aufbau lediglich für eine wesentlich kleinere Anzahl und zufällig ausgewählter Phagensequenzen gezeigt (Kwan *et al.* 2005, Goerke *et al.* 2009).

6.2.3 Prophagen häufen kontinuierlich Sequenzunterschiede an

Dynamiken der Diversifizierung von Prophagen. Die große Diversität von Prophagen im Allgemeinen (Brüssow, Canchaya & Hardt 2004), in *S. aureus* (Kwan *et al.* 2005, Goerke *et al.* 2009) und für den klonalen Komplex CC5 (diese Arbeit) wurde bereits gezeigt. Die Frage, welche Mechanismen zu dieser Variabilität geführt haben und wie häufig diese auftreten, konnte bisher jedoch nicht beantwortet werden.
Für diese Arbeit wurden die Genome von 24 *S. aureus* Isolaten des klonalen Komplexes CC5 mit „*next generation*“-Sequenziertechnologien sequenziert. Damit ist die Voraussetzung für zwei notwendige Informationen gegeben, um die oben gestellte Frage beantworten zu können: Kenntnisse über die Diversität der Prophagen und die hochaufgelöste Phylogenie der bakteriellen Wirte.
Wie bereits oben beschrieben, wurde das Programm ClonalFrame (Didelot & Falush 2007) genutzt, um in den Prophagen-Sequenzen Rekombinationsereignisse zu detektierten. Diese Ereignisse wurden anschließend auf die Bakterien-Phylogenie abgebildet. So kann das Auftreten von Rekombination zeitlich eingeordnet werden. Die vorliegenden Ergebnisse zeigen, dass die Diversität von Prophagen im klonalen Komplex CC5

vor allem durch den häufigen Austausch von DNA-Fragmenten durch Rekombination entstanden ist.
Die Sa3*int*-Prophagen machen den größten Anteil der Prophagen im klonalen Komplex CC5 aus. Es ist davon auszugehen, dass die Ergebnisse der Analysen robust sind, weshalb nur auf diese näher eingegangen wird. Die Abbildung der Rekombinationsereignisse auf die Phylogenie zeigt, dass genetischer Austausch sowohl kürzlich (Spitzen im Stammbaum) als auch schon früher (innen liegende, ältere Äste) stattgefunden hat (Abbildung 5.11A). Dieses Ergebnis ist konträr zu Beschreibungen von Prophagen in *Streptococcus pyogenes*, deren Diversifizierung eher ein junges Ereignis ist (Marri, Hao & Golding 2006, Didelot, Darling & Falush 2009).
Die Diversifizierung von Prophagen des klonalen Komplexes CC5 ist also offenbar ein fortwährender Prozess. Trägt man die Anzahl der Rekombinationereignisse in den Prophagen gegen die paarweise Distanz der bakteriellen Wirte auf, erhält man eine signifikante Korrelation ($p = 0{,}001$; Abbildung 5.11B). Es ist auch zu sehen, dass ab einer gewissen Diversität der Wirtsbakterien (Distanz ca. 130 SNPs im Bakterien-Kerngenom) die Auswirkung der Rekombination nicht mehr zunimmt. Am wahrscheinlichsten ist es, dass eine Sättigung der Sequenzen mit importierten Fragmenten vorliegt, die eine Detektierung zusätzlicher Ereignisse verhindert.
Zusammenfassend lässt sich sagen, dass Sequenzunterschiede über die Zeit angehäuft werden und eine kontinuierliche Umgestaltung des Prophagen-Genoms durch homologe Rekombination als antreibende Kraft beobachtet werden kann.

Bemerkungen zu den angewendeten Methoden. Wie bereits in den bioinformatischen Methoden beschrieben (Kapitel 4.11), ist ein gutes Alignment essentiell für die verlässliche Identifizierung von Rekombinationsereignissen. Die große Diversität innerhalb der Prophagen erschwert die Alignierung der Sequenzen.
Es ist allerdings bekannt, dass trotz großer Sequenzunterschiede auf Nukleotidebene die Gensequenz auf Aminosäurelevel Übereinstimmungen zeigen kann. Um ein Alignment zu verbessern, wäre es also sinnvoll eine Software zu verwenden, die intergenische Regionen auf Nukleotidebene und kodierende Bereiche auf Aminosäurelevel aligniert.
So kommt es nicht zur Detektion von falsch positiven Sequenzunterschieden und damit zu einer Überschätzung der Anzahl von Rekombinationsereignissen. Leider lag solch ein Programm zum Zeitpunkt der Erstellung dieser Arbeit nicht vor. Für die vorliegenden Daten wurden daher zu diverse Phagensequenzen von den Analysen ausgeschlossen, um verlässliche Ergebnisse sicherzustellen.

6.3 Mikroevolution des Sequenztyps ST225

6.3.1 Phylogenetische Methoden geeignet zur Aufklärung von Transmissionswegen

Seit dem Aufkommen der ersten Sequenziertechnologien ist es möglich, phylogenetische Fragestellungen auch ohne morphologische Merkmale zu beantworten. Meilensteine waren z.B. die Einteilung der Lebewesen in drei Domänen basierend auf hochkonservierten 16S rRNA-Sequenzen (Woese, Kandler & Wheelis 1990) sowie die Aufklärung des Ursprungs des HI-Virus bei den Menschenaffen Westafrikas (Keele *et al.* 2006). Transmission des HI-Virus zwischen Menschen wurde mittels phylogenetischer Methoden bereits in vielen Fällen erfolgreich aufgedeckt (Ou *et al.* 1992, Hillis & Huelsenbeck 1994, Blanchard *et al.* 1998, Goujon *et al.* 2000).
Im Folgenden soll gezeigt werden, inwieweit ganze Genomsequenzen in Kombination mit phylogenetischen Methoden geeignet sind, um die Transmission von MRSA - hier am Beispiel des Sequenztyps ST225 - nachverfolgen zu können. Dazu wurden zwei Datensätze untersucht: Ersterer umfasst eine globale Sammlung von Isolaten und der zweite besteht aus Isolaten aus einem dänischen Krankenhaus.

Kontinuierliche Anhäufung von SNPs. Die Verwendung ganzer Genomsequenzen bestätigt die geringe Diversität im Sequenztyp ST225, welche bereits von Nübel *et al.* (2010) beschrieben wurde.
Die durchgeführten linearen Regressionsanalysen unter Berücksichtigung des Isolationsdatums und der Distanz zwischen Wurzel und terminalen Ästen („*root-to-tip*"-Distanz; Kapitel 5.4.7 *Sequenzvariationen* und Abbildung 5.18) zeigen trotzdem eine kontinuierliche Anhäufung von SNPs. Diese ist in den dänischen Patientenisolaten sogar noch stärker ausgeprägt als im globalen Datensatz. Es ist davon auszugehen, dass Sequenzvariationen während der Kolonisierung bzw. der Infektion mit *S. aureus* entstehen.
Die Anhäufung von SNPs und die signifikante Korrelation zwischen „*root-to-tip*"-Distanz und Isolationsdatum innerhalb eines Ausbruchszeitraum sind notwendige Voraussetzungen für die Zurückverfolgung von Transmissionswegen. Dadurch ist es möglich, sowohl die zeitliche Abfolge als auch die Richtung der Übertragung zu rekonstruieren. Letzteres ist mit herkömmlichen Typisierungsmethoden bisher nicht durchführbar gewesen.

Interkontinentale und internationale Transmissionen. Nübel *et al.* (2010) konnten mit ihren Analysen zur räumlichen Ausbreitung des Sequenztyps ST225 zeigen, dass Methicillin-resistente *S. aureus* auch in relativ kurzer Zeit genug Sequenzunterschiede anhäufen, um deren Ausbreitung über lange Distanzen wie z.B. zwischen Kontinenten oder Ländern rekonstruieren zu können. Die vorliegenden phylogeneti-

schen Rekonstruktionen unter Verwendung ganzer Genomsequenzen bestätigen dies (Abbildung 5.14). So spricht die basale Stellung der US-amerikanischen Isolate für einen nordamerikanischen Ursprung des Sequenztyps ST225. Die basale Stellung der deutschen Isolate innerhalb der europäischen Isolate zeigt, dass sich der Sequenztyp ST225 von Deutschland ausgehend in Europa verbreitet hat (Abbildung 5.14).
Die aus den zwei vorliegenden Datensätzen rekonstruierten Stammbäume zeigen deutlich zwei unabhängige Ausbruchsgeschehen des Sequenztyps ST225 in einem Krankenhaus in Kopenhagen, Dänemark (Abbildung 5.15). Aufgrund der zuvor verwendeten, konventionellen Typisierungsmethoden war es nicht möglich, diese Ausbrüche zu unterscheiden. Mitarbeiter des Krankenhauses identifizierten in ihren Untersuchungen eine Index-Patientin aus Frankfurt am Main, Deutschland mit der 2004 ein Klon des Sequenztyps ST225 in das dänische Krankenhaus eingeschleppt wurde. Das Isolat der Index-Patientin nimmt die basale Stellung im zugehörigen Ausbruchs-Cluster der vorliegenden Stammbäume ein (Abbildung 5.15). Die Genomsequenz des Isolats bestätigt damit den Verdacht der Krankenhaus-Mitarbeiter.

Identifikation von Infektionsquellen und Transmissionswegen. Harris *et al.* (2010) fanden in *S. aureus*-Genomsequenzen des Sequenztyps ST239 so viel Variabilität in Isolaten aus einem thailändischen Krankenhauses, dass ihrer Vermutung nach Genomsequenzen genug Auflösungsvermögen bieten könnten, um die Transmission zwischen oder innerhalb eines Krankenhauses aufzuklären. Für die Nachverfolgung der räumlichen Ausbreitung von MRSA auf einer globalen, nationalen und lokalen Ebene sind phylogenetische Methoden also wahrscheinlich gut geeignet. Unklar ist bisher jedoch, ob dies auch für die Aufklärung von Transmissionsketten zwischen einzelnen Patienten z.B. während eines Ausbruchs in einem Krankhaus möglich ist.

Wie weiter oben bereits beschrieben, wurden phylogenetische Methoden bereits erfolgreich zur Aufklärung von HIV-Transmisson zwischen Menschen angewendet. Scaduto *et al.* (2010) konnten aber sogar die Richtung von zwei HIV-Infektionen (Quelle → Empfänger) aufklären.
Dabei machten sie sich das Verwandtschaftsverhältnis der Paraphylie zu Nutzen, das in phylogenetischen Stammbäumen beobachtet werden kann. Im Gegensatz zu monophyletischen Clustern, die alle von einer gemeinsamen Stammform abgeleiteten Taxa enthalten, versteht man unter paraphyletischen Clustern, Taxagruppen, die trotz gemeinsamer Stammform nicht alle Taxa einschließen.[5]
Shattock & Moore (2003) verstehen Paraphylie bei HIV als Ergebniss eines genetischen Flaschenhalses, der bei der Etablierung einer HIV-Infektion in einem Patienten auf-

[5]Paraphylie: Ein gutes Beispiel für eine paraphyletische Gruppe sind die Reptilien. Das Paraphylum der Reptilien schließt die Vögel aus, obwohl beide Taxa auf eine gemeinsame Stammform zurück gehen.

tritt. Sequenzen, die bezüglich weiterer untersuchter Sequenzen paraphyletisch sind, repräsentieren nach den Ergebnissen von Scaduto *et al.* (2010) die Quelle der Infektion, also den Index-Patienten. Diese Analysen hatten juristische Relevanz: In zwei Fällen konnten sie die Täter überführen, die mehrere Frauen mit HIV infiziert hatten, was zur Verurteilung der beiden Männer führte. Im Folgenden soll gezeigt werden, dass für MRSA ähnliche Methoden wie für schnell evolvierende Viren angewendet werden können.

Der „*Maximum Likelihood*"-Baum der Patientenisolate (Abbildung 5.16) zeigt in einigen Fällen ebenfalls Paraphylie. Hier soll besonders auf das Cluster des Patienten 4 eingegangen werden.
Das Cluster des Patienten 4 ist paraphyletisch, da es auch das Isolat 07-03032 enthält. Die Paraphylie der Patient 4-Isolate gegenüber 07-03032 wird mit einem approximierten Bootstrap-Wert von 98 % signifikant unterstützt. Da der letzte gemeinsame Vorfahre ein Isolat des Patienten 4 ist, ist die Richtung der Transmisson von Patient 4 auf den Patienten des Isolats 07-03032. Beide Patienten lagen gleichzeitig in der geriatrischen Abteilung des Krankenhauses, und die Besiedlung des Patienten 4 war zu diesem Zeitpunkt bereits bekannt.

Abschließende Bemerkungen. Obwohl weitere Analysen mit mehr Isolaten und mehr epidemiologischen Daten zur Bestätigung der vorliegenden Ergebnisse durchgeführt werden sollten, konnte gezeigt werden, dass die Verwendung ganzer Genomsequenzen in Kombination mit phylogenetischen Methoden ein vielversprechender Ansatz zur Aufklärung der räumlichen Ausbreitung von MRSA ist.
Die Transmission des Sequenztyps ST225 konnte auf verschiedenen Ebenen rekonstruiert werden: lange Distanzen (interkontinental), kurze Distanzen (international) und lokale Ausbrüche. Außerdem konnte mit den vorliegenden Daten in einem Fall sogar die wahrscheinliche Richtung der Übertragung zwischen zwei Patienten rekonstruiert werden.
Diese Ergebnisse haben große Bedeutung für die Entwicklung neuer Typisiermethoden mit hoher Diskriminierungsfähigkeit. Die Aufdeckung von Ausbruchsgeschehen, die mit aktuell verwendeten Typisierungsmethoden nicht unterschieden werden können, sowie Kenntnisse über die Wege und Richtungen der Ausbreitung können helfen, eine weitere Verbreitung einzudämmen. So können sowohl schwerwiegende Erkrankungen von Menschen als auch immense Kosten für das Gesundheitssystem vermieden werden.

6.3.2 Untersuchungen zur Evolution von ST225 in Patienten

Die Evolution von MRSA des klonalen Komplexes CC5 wurde in dieser Arbeit auf unterschiedlichen Ebenen untersucht. Angefangen von Vergleichen zwischen globalen

und regionalen Isolaten zu Isolaten eines Krankenhauses soll nun auf die Evolution von Isolaten in einzelnen Patienten eingegangen werden.

Trotz ansteigender Distanz keine Zunahme der Diversität in Patientenisolaten. Im Rahmen dieser Arbeit wurden aus fünf Patienten je eine Reihe von Stämmen isoliert und sequenziert. Damit kann die Evolution von MRSA innerhalb einzelner Patienten gezeigt werden. Dies ist möglich, da die Genome kontinuierlich SNPs anhäufen (Kapitel 5.4.6 *Sequenzvariationen* und Abbildung 5.18).
Für vier der fünf Patienten wurden phylogenetische Bäume erstellt. Angegeben sind das Isolationsdatum und die Entnahmestelle für jedes Isolat sowie für jeden Patienten die verabreichten Therapeutika (Abbildung 5.17).
Die zeitgleiche Entnahme von Isolaten an verschiedenen Körperstellen zeigt keine bis wenig Diversität. Obwohl sich die Isolate also mit der Zeit genetisch von dem „Gründer-Genom“ entfernen und SNPs akkumulieren, nimmt die Diversität nicht zu. Dieses Phänomen wurde auch für *Pseudomonas aeruginosa* in Patienten mit Mukoviszidose beschrieben (Yang *et al.* 2011). Für *S. aureus* ist dies vermutlich damit zu begründen, dass ein Klon mutiert und sich dann über den Körper hinweg ausbreitet.
Drei der vier näher untersuchten Patienten wurden u.a. mit Antibiotika behandelt. Allerdings haben sich in keinem der untersuchten Stämme neue Resistenzen ausgebildet. Das beschriebene evolutionäre Muster (keine Zunahme von Diversität) wird auch für den Patienten ohne Behandlung beobachtet (Patient 4). Die neue und sich ausbreitende Variante muss also nicht notwendigerweise einen selektiven Vorteil haben. Die Berechnung der Evolutionsraten für Patienten-Isolate und verschiedene Sequenztypen zeigt ähnliche Raten: Die Gabe von Therapeutika scheint also keinen Einfluss auf die Geschwindigkeit der MRSA-Evolution in Patienten zu haben.

Einfluss von Selektion. Populationsweit liegt $\mathrm{d}N/\mathrm{d}S$ für den Sequenztyp ST225 bei 0,6, was auf einen mäßigen Einfluss reinigender Selektion hindeutet. Interessanter ist $\mathrm{d}N/\mathrm{d}S$ für den Patienten-Datensatz: $\mathrm{d}N/\mathrm{d}S$ ist am Anfang sehr hoch (1,5 - 1,3) und nimmt dann sehr schnell ab und pendelt sich - wie schon populationsweit beobachtet - auf 0,6 ein (Abbildung 5.19). Der hohe Anfangswert ist mit der gerade beginnenden Diversifizierung der verglichenen Genome zu erklären. Anzumerken ist, dass solch ein hoher Wert für *S. aureus* bisher nicht beschrieben wurde.
Zusätzlich zu $\mathrm{d}N/\mathrm{d}S$-Analysen wurde auch die Häufigkeit der Entstehung verschiedener Mutationsmuster in den Patienten-Isolaten untersucht. Hershberg & Petrov (2010) und Hildebrand, Meyer & Eyre-Walker (2010) fanden unabhängig voneinander eine generelle Tendenz zu AT-anreichernden Mutationen. Dies würde mit der Zeit jedoch zu einem geringen GC-Gehalt in allen bakteriellen Genomen führen. Die vorliegenden Analysen zeigen einen Überschuss an AT-Mutationen in sehr nah verwandten Isolaten

(Abbildung 5.19). Der Anteil an AT-Mutationen nimmt jedoch mit ansteigender Distanz zwischen zwei Stämmen ab und ist so konsistent mit dem Verlauf von dN/dS. Die Anreicherung des Genoms mit AT-Mutationen ist also ein sehr junges Ereigniss und Mutationen in den weiter innenliegenden - älteren - Ästen wurden bereits entfernt. Dadurch bleibt der GC-Gehalt des Genoms über längere Zeiträume konstant. Die wirkenden Mechanismen sind bisher allerdings unbekannt.
In letzter Zeit haben sich einige Publikationen mit genau diesem Thema beschäftigt, was die Aktualität der vorliegenden Analysen zeigt (Balbi *et al.* 2009, Hershberg & Petrov 2010, Hildebrand, Meyer & Eyre-Walker 2010, Rocha & Feil 2010).

6.4 „*Next Generation Sequencing*“: Ausblick, Anwendungen, Limitierungen

Im Folgenden soll als Abschluss dieser Arbeit auf die Möglichkeiten und die vorhandenen Limitierungen des Einsatzes von „*next generation sequencing*“-Technologien eingegangen werden.

Kenntnisse über die Verwandtschaft von pathogenen Erregern, die Beobachtung ihrer Ausbreitung sowie die Aufklärung von Infektionsketten ist ein wichtiger Bestandteil der Überwachung von Infektionskrankheiten. Die Einführung von Typisierungsmethoden war daher ein Meilenstein der epidemiologischen Forschung. Es kann zwischen phäno- und genotypischen Techniken unterschieden werden, die je nach Fragestellung Vor- und Nachteile haben. Eine sequenzbasierte, genotypische Methode ist die Multilokus Sequenztypisierung (MLST), die auf Variationen innerhalb weniger Gene beruht. Obwohl diese Methode wichtige Einblicke in die Populationsstruktur und Evolution verschiedener Bakterien geben konnte, ist sie aufgrund der geringen Diskrimination für Kurzzeitepidemiologie nicht aussagekräftig. Für diese müssen Methoden mit einer höheren Auflösung verwendet werden, wie sie z.B. die Verwendung ganzer Genomsequenzen liefert.

Die Sequenzierung ganzer Genome hat in den letzten Jahren stark zugenommen, was an der schnellen Entwicklung neuer und kostengünstiger Sequenziertechnologien liegt. Sequenzen ganzer Genome ermöglichen die Rekonstruktion robuster Phylogenien, die für die Identifizierung von Infektionsquellen sowie zur Aufklärung von Transmissionswegen verwendet werden können.
Diese Arbeit begann mit der Rekonstruktion der globalen Phylogenien des klonalen Komplexes CC5 und des Sequenztyps ST225 und ging in eine Aufklärung der Ausbreitung des Sequenztyps ST225 über internationale und nationale Grenzen über. Außerdem konnten zwei unterschiedliche Ausbruchsgeschehen des gleichen

Sequenztyps (ST225) in einem dänischen Krankenhaus voneinander unterschieden werden. Harris *et al.* (2010) vermuteten, dass die Verwendung ganzer Genomsequenzen genug Informationen für eine Rekonstruktion von Transmissionswegen zwischen verschiedenen Abteilungen eines Krankenhauses enthalten könnte. In dieser Arbeit konnte zusätzlich in einigen Fällen die Verbreitung von MRSA zwischen sowie die Evolution innerhalb einzelner Patienten gezeigt werden.

In Zukunft wird die Sequenzierung bakterieller Genome immer schneller und günstiger werden. Mit der Einführung von sogenannten „*benchtop*"-Geräten (z.B. GS Junior 454/Roche, Ion Torrent PGM oder Illumina MiSeq) können sich schon bald auch kleine Labore die Sequenzierung ganzer Genome leisten und die Sequenzierung ganzer Genome wird eine wichtige Rolle in der mikrobiellen Diagnostik spielen.
Trotz aller Vorteile gibt es noch einige nicht unerhebliche Limitierungen. Die erste Schwachstelle ist die Vorbereitung der Proben (DNA-Extraktion, Erstellung der Sequenzier-Bibliotheken), die - für einen Einsatz in der Diagnostik - noch immer recht langwierig und schwierig ist. Ein größeres Problem ist aber die mit den Sequenzierungsprozessen einhergehende Generierung von immensen Datenmengen. Diese stellt Informationstechnologie und Bioinformatik vor neue Herausforderungen, was die Speicherung, Verwaltung und Analyse sowie die biologische und klinische Interpretation der Daten angeht (Su *et al.* 2011).
Parkhill & Wren (2011) sehen im „*next generation sequencing*" eine neue Ära der mikrobiellen Genomik, durch die die historische und geographische Ausbreitung von Bakterien rekonstruiert werden kann und die unser Verständnis bakterieller Evolution grundlegend verändern wird. Damit geht auch eine Umgestaltung der molekularen Epidemiologie von klonalen bakteriellen Pathogenen einher. Beispielsweise ist ein Ausbruchs-Frühwarnsystem denkbar, bei dem die Genom-Daten mit Geo-Informationssystemen und Raum-Zeit-Clusteranalysen kombiniert werden. Abschließen möchte ich meine Arbeit mit einem Zitat von Parkhill & Wrenn (2011):

> *The next few years promise a voyage of discovery in terms of the attribution of sources and transmission tracking of bacteria, the understanding of how and why epidemic clones emerge or dissapear, and ultimately the managment and treatment of infectious diseases.*

Literaturverzeichnis

Abu-Qatouseh LF, Chinni SV, Seggewiss J, Proctor RA, Brosius J, Rozhdestvensky TS, Peters G, von Eiff C & Becker K (2010). Identification of differentially expressed small non-protein-coding RNAs in *Staphylococcus aureus* displaying both the normal and the small-colony variant phenotype. *J Mol Med* **88**(6): 565-575.

Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, Kusecek B, Vogler AJ, Wagner DM, Allender CJ, Easterday WR, Chenal-Francisque V, Worsham P, Thomson NR, Parkhill J, Lindler LE, Carniel E & Keim P (2004). Microevolution and history of the plague bacillus, *Yersinia pestis*. *Proc Natl Acad Sci USA* **101**(51): 17837-17842.

Aires de Sousa M, Correia B, de Lencastre H & the Multilaboratory Project Collaborators (2008). Changing Patterns in Frequency of Recovery of Five Methicillin-Resistant *Staphylococcus aureus* Clones in Portuguese Hospitals: Surveillance over a 16-Year Period. *J Clin Microbiol* **46**(9): 2912-2917.

Akaike H (1974). A new look at the statistical model identification. *IEEE Trans Autom Control* **19**(6): 716-723.

Aldeyab MA, Monnet DL, López-Lozano JM, Hughes CM, Scott MG, Kearney MP, Magee FA & McElnay JC (2008). Modelling the impact of antibiotic use and infection control practices on the incidence of hospital-acquired methicillin-resistant *Staphylococcus aureus*: a time-series analysis. *J Antimicrob Chemother* **62**(3): 593-600.

Alland D, Whittam TS, Murray MB, Cave MD, Hazbon MH, Dix K, Kokoris M, Duesterhoeft A, Eisen JA, Fraser CM & Fleischmann RD (2003). Modeling bacterial evolution with comparative-genome-based marker systems: application to Mycobacterium tuberculosis evolution and pathogenesis. *J Bacteriol* **185**(11): 3392-3399.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol* **215**(3): 403-410.

Assefa S, Keane TM, Otto TD, Newbold C & Berriman M (2009). ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**(15): 1968-1969.

Baker L, Brown T, Maiden MC & Drobniewski F (2004). Silent nucleotide polymorphisms and a phylogeny for *Mycobacterium tuberculosis*. *Emerg Infect Dis* **10**(9): 1568-1577.

Balbi KJ, Rocha EP & Feil EJ (2008). The temporal dynamics of slightly deleterious mutations in *Escherichia coli*and *Shigella* spp. *Mol Biol Evol* **26**(2): 345-355.

Barksdale L & Arden SB (1974). Persisting bacteriophage infections, lysogeny, and phage conversions. *Annu Rev Microbiol* **28**(0): 265-299.

Beres SB, Carroll RK, Shea PR, Sitkiewicz I, Martinez-Gutierrez JC, Low DE, McGeer A, Willey BM, Green K, Tyrrell GJ, Goldman TD, Feldgarden M, Birren BW, Fofanov Y, Boos J, Wheaton WD, Honisch C & Musser JM (2010). Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. *Proc Natl Acad Sci USA* **107**(9): 4371-4376.

Blanchard A, Ferris S, Chamaret S, Guetard D & Montagnier L (1998). Molecular evidence for nosocomial transmission of human immunodeficiency virus from a surgeon to one of his patients. *J Virol* **72**(5): 4537-4540.

Bohn C, Rigoulay C, Chabelskaya S, Sharma CM, Marchais A, Skorski P, Borezee-Durant E, Barbet R, Jacquet E, Jacq A, Gautheret D, Felden B, Vogel J & Bouloc P (2010). Experimental discovery of small RNAs in *Staphylococcus aureus* reveals a riboregulator of central metabolism. *Nucleic Acids Res* **38**(19): 6620-6636.

Bonnet E & Van de Peer Y (2002). zt: A Sofware Tool for Simple and Partial Mantel Tests. *J Stat Soft* **7**(10): 1-12.

Brüssow H, Canchaya C & Hardt WD (2004). Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol Mol Biol Rev* **68**(3): 560-602.

Brüssow H & Desiere F (2001). Comparative phage genomics and the evolution of *Siphoviridae*: insights from dairy phages. *Mol Microbiol* **39**(2): 213-222.

Campanile F, Bongiorno D, Borbone S & Stefani S (2009). Hospital-associated methicillin-resistant *Staphylococcus aureus* (HA-MRSA) in Italy. *Ann Clin Microbiol Antimicrob* **8**(22).

Castillo-Ramírez S, Harris SR, Holden MTG, He M, Parkhill J, Bentley SD & Feil EJ (2011). The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathog* **7**(7): e1002129.

Canchaya C, Proux C, Fournous G, Bruttin A & Brussow H (2003). Prophage genomics. *Microbiol Mol Biol Rev* **67**(2): 238-276.

Chambers HF (2001). The changing epidemiology of *Staphylococcus aureus*? *Emerg Infect Dis* **7**(2): 178-182.

Chambers HF & DeLeo FR (2009). Waves of resistance: *Staphylococcus aureus* in the antibiotic era. *Nat Rev Microbiol* **7**(9): 629-641.

Chibani-Chennoufi S, Sidoti J, Bruttin A, Kutter E, Sarker S & Brussow H (2004). In vitro and in vivo bacteriolytic activities of *Escherichia coli* phages: implications for phage therapy. *Antimicrob Agents Chemother* **48**(7): 2558-2569.

Chongtrakool P, Ito T, Ma XX, Kondo Y, Trakulsomboon S, Tiensasitorn C, Jamklang M, Chavalit T, Song JH & Hiramatsu K (2006). Staphylococcal cassette chromosome *mec* (SCC*mec*) typing of methicillin-resistant *Staphylococcus aureus* strains isolated in 11 Asian countries: a proposal for a new nomenclature for SCC*mec* elements. *Antimicrob Agents Chemother* **50**(3): 1001-1012.

Clark AJ, Inwood W, Cloutier T & Dhillon TS (2001). Nucleotide sequence of coliphage HK620 and the evolution of lambdoid phages. *J Mol Biol* **311**(4): 657-679.

Cock PJA, Fields CJ, Goto N, Heuer ML & Rice PM (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**(6): 1767-1771.

Coleman D, Knights J, Russell R, Shanley D, Birkbeck TH, Dougan G & Charles I (1991). Insertional inactivation of the *Staphylococcus aureus* beta-toxin by bacteriophage phi 13 occurs by site- and orientation-specific integration of the phi 13 genome. *Mol Microbiol* **5**(4): 933-939.

Conceicao T, Aires de Sousa M, Füzi M, Tóth A, Pászti J, Ungvari E, Van Leeuwen WB, Van Belkum A, Grundmann H & de Lencastre H (2007). Replacement of methicillin-resistant *Staphylococcus aureus* clones in Hungary over time: a 10-year surveillance study. *Clin Microbiol Infect* **13**(10): 971-979.

Cooper BS, Medley GF, Stone SP, Kibbler CC, Cookson BD, Roberts JA, Duckworth G, Lai R & Ebrahim S (2004). Methicillin-resistant *Staphylococcus aureus* in hospitals and the community: stealth dynamics and control catastrophes. *Proc Natl Acad Sci USA* **101**(27): 10223-10228.

Crisóstomo MI, Westh H, Tomasz A, Chung M, Oliveira DC & de Lencastre H (2001). The evolution of methicillin resistance in *Staphylococcus aureus*: Similarity of genetic backgrounds in historically early methicillin-susceptible and -resistant isolates and contemporary epidemic clones. *Proc Natl Acad Sci USA* **98**(17): 9865-9870.

Cuny C & Layer F (2011). Auftreten und Verbreitung von MRSA in Deutschland 2010 *in* Epidemiologisches Bulletin **26/2011**. Robert Koch-Institut.

Darling AE, Mau B & Perna NT (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *Bioinformatics* **5**(6): e11147.

Deurenberg RH, Vink C, Kalenic S, Friedrich AW, Bruggeman CA, Bruggeman CA & Stobberingh EE (2007). The molecular evolution of methicillin-resistant *Staphylococcus aureus*. *Clin Microbiol Infect* **13**(3): 222-235.

Didelot X, Barker M, Falush D & Priest FG (2009). Evolution of pathogenicity in the *Bacillus cereus* group. *Syst Appl Microbiol* **32**(2): 81-90.

Didelot X, Darling A & Falush D (2009). Inferring genomic flux in bacteria. *Genome Res* **19**(2): 306-317.

Didelot X & Falush D (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**(3): 1251-1266.

Drummond AJ & Rambaut A (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**(214).

Drummond AJ, Rambaut A & Suchard M (2010). „BEAUTi v1.6.1.“ *erhältlich unter* `http://beast.bio.ed.ac.uk/BEAUti` [Stand: 03.08.2011].

Edgeworth JD (2011). Has decolonization played a central role in the decline in UK methicillin-resistant *Staphylococcus aureus* transmission? A focus on evidence from intensive care. **J Antimicrob Chemother 66** Suppl 2: 41-47.

Enright MC, Robinson DA, Randle G, Feil EJ, Grundmann H & Spratt BG (2002). The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). *Proc Natl Acad Sci USA* **99**(11): 7687-7692.

Enrigth AJ, Van Dongen S & Ouzounis CA (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**(7): 1575-1584.

Eriksen KR & Erichesen I (1964). Resistance to Methicillin, Isoxazolyl Penicillins, and Cephalothin in *Staphylococcus aureus*. *Acta Pathol Microbiol Scand* **52**: 255-275.

Ewing B & Green P (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**(3): 186-194.

Ewing G, Hillier L, Wendl MC & Green P (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**(3): 178-185.

Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, Primeaux C, Sezate S, Suvorov AN, Kenton S, Lai HS, Lin SP, Qian Y, Jia HG, Najar FZ, Ren Q, Zhu H, Song L, White J, Yuan X, Clifton SW, Roe BA & McLaughlin R (2001). Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc Natl Acad Sci USA* **98**(8): 4658-4663.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb J-F, Dougherty BA, Merrick JM, McKenney K, Sutton G, Fitzhugh W, Fields C, Gocyne JD, Scott J, Shirley R, Liu L-I, Glodek A, Kelley JM, Weidman JF, Phillips CA, Spriggs T, Hedblom E, Cotton MD, Utterback TR, Hanna MC, Nguyen DT, Saudek DM, Brandon RC, Fine LD, Fritchman JL, Fuhrmann JL, Geoghagen NSM, Gnehm CL, McDonald LA, Small KV, Fraser CM, Smith HO & Venter JC (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**(5223): 496-512.

Freeman VJ (1951). Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *J Bacteriol* **61**(6): 675-688.

Geissmann T, Chevalier C, Cros MJ, Boisset S, Fechter P, Noirot C, Schrenzel J, Francois P, Vandenesch F, Gaspin C & Romby P (2009). A search for small noncoding RNAs in *Staphylococcus aureus* reveals a conserved sequence motif for regulation. *Nucleic Acids Res* **37**(21): 7239-7257.

Gillet Y, Issartel B, Vanhems P, Fournet JC, Lina G, Bes M, Vandenesch F, Piémont Y, Brousse N, Floret D & Etienne J (2002). Association between *Staphylococcus aureus* strains carrying gene for Panton-Valentine leukocidin and highly lethal necrotising pneumonia in young immunocompetent patients. *Lancet* **359**(9308): 753-759.

Goerke C, Pantucek R, Holtfreter S, Schulte B, Zink M, Grumann D, Broker BM, Doskar J & Wolz C (2009). Diversity of prophages in dominant *Staphylococcus aureus* clonal lineages. *J Bacteriol* **191**(11): 3462-3468.

Gojobori T, Ishii K & Nei M (1982). Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *J Mol Evol* **18**(6): 414-423.

Gomes AR, Westh H & de Lencastre H (2006). Origins and evolution of methicillin-resistant *Staphylococcus aureus* clonal lineages. *Antimicrob Agents Chemother* **50**(10): 3237-3244.

Goujon CP, Schneider VM, Grofti J, Montigny J, Jeantils V, Astagneau P, Rozenbaum W, Lot F, Frocrain-Herchkovitch C, Delphin N, Le Gal F, Nicolas JC, Milinkovitch MC & Deny P (2000). Phylogenetic analyses indicate an atypical nurse-to-patient transmission of human immunodeficiency virus type 1. *J Virol* **74**(6): 2525-2532.

Gray RR, Tatem AJ, Johnson JA, Alekseyenko AV, Pybus OG, Suchard MA & Salemi M (2011). Testing spatiotemporal hypothesis of bacterial evolution using methicillin-resistant *Staphylococcus aureus* ST239 genome-wide data within a bayesian framework. *Mol Biol Evol* **28**(5): 1593-1603.

Gupta S, Ferguson N & Anderson R (1998). Chaos, persistence, and evolution of strain structure in antigenically diverse infectious agents. *Science* **280**(5365): 912-915.

Harris SR, Feil EJ, Holden MT, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ & Bentley SD (2010). Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**(5964): 469-474.

Hartmann AA (1978). Staphylococci of the normal human skin flora. Variety of biotypes and antibiograms without direct correlations. *Arch Dermatol Res* **261**(3): 295-302.

Hatfull GF (2008). Bacteriophages: nature's most successful experiment. *Microbiology Today*: 188-191.

Hatfull GF (2010). Mycobacteriophages: genes and genomes. *Annu Rev Microbiol* **64**: 331-356.

Hendrix RW, Smith MC, Burns RN, Ford ME & Hatfull GF (1999). Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci USA* **96**(5): 2192-2197.

Hershberg R & Petrov DA (2010). Evidence that mutation is universally biased towards AT in bacteria *PLoS Genet* **6**(9): e1001115.

Hildebrandt F, Meyer A & Eyre-Walker A (2010). Evidence of selection upon genomic GC-content in bacteria *PLoS Genet* **6**(9): e1001107.

Hillis DM & Huelsenbeck JP (1994). Support for dental HIV transmission. *Nature* **369**(6475): 24-24.

Hiramatsu K, Cui L, Kuroda M & Ito T (2001). The emergence and evolution of methicillin-resistant *Staphylococcus aureus*. *Trends Microbiol* **9**(10): 486-493.

Holden MT, Feil EJ, Lindsay JA, Peacock SJ, Day NP, Enright MC, Foster TJ, Moore CE, Hurst L, Atkin R, Barron A, Bason N, Bentley SD, Chillingworth C, Chillingworth T, Churcher C, Clark L, Corton C, Cronin A, Doggett J, Dowd L, Feltwell T, Hance Z, Harris B, Hauser H, Holroyd S, Jagels K, James KD, Lennard N, Line A, Mayes R, Moule S, Mungall K, Ormond D, Quail MA, Rabbinowitsch E, Rutherford K, Sanders M, Sharp S, Simmonds M, Stevens K, Whitehead S, Barrell BG, Spratt BG & Parkhill J (2004). Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci USA* **101**(26): 9786-9791.

Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, Dolecek C, Achtman M & Dougan G (2008). High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi. *Nat Genet* **40**(8): 987-993.

Huelsenbeck JP & Ronquist F (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**(8): 754-755.

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M & Rupp R (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**(460).

Iandolo JJ, Worrell V, Groicher KH, Qian Y, Tian R, Kenton S, Dorman A, Ji H, Lin S, Loh P, Qi S, Zhu H & Roe BA (2002). Comparative analysis of the genomes of the temperate bacteriophages phi 11, phi 12 and phi 13 of *Staphylococcus aureus* 8325. *Gene* **289**(1-2): 109-118.

Illumina Inc. (2011). History of Solexa Sequencing. Online Publikation `http://www.illumina.com/technology/solexa_technology.ilmn` [Stand: 12. Juli 2011].

International Working Group on the Staphylococcal Cassette Chromosome elements (2011). Currently identified SCC*mec* types in *S. aureus* strains. Online Publikation `http://www.sccmec.org/Pages/SCC_TypesEN.html` [Stand: 20.10.2011].

Iwase T, Uehara Y, Shinji H, Tajima A, Seo H, Takada K, Agata T & Mizunoe Y (2010). *Staphylococcus epidermidis* Esp inhibits *Staphylococcus aureus* biofilm formation and nasal colonization. *Nature* **465**(7296): 346-349.

Jevons MP (1961). „Celbenin - resistant Staphylococci. *Br Med J* **1**(5219): 124-125.

Jikia D, Chkhaidze N, Imedashvili E, Mgaloblishvili I, Tsitlanadze G, Katsarava R, Glenn Morris J & Sulakvelidze A (2005). The use of a novel biodegradable preparation capable of the sustained release of bacteriophages and ciprofloxacin, in the complex treatment of multidrug-resistant *Staphylococcus aureus*-infected local radiation injuries caused by exposure to Sr90. *Clin Exp Dermatol* **30**(1): 23-26.

Jobb G, von Haeseler A & Strimmer K (2004). TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* **4**(18).

Johnsson D, Mölling P, Strålin K & Söderquist B (2004). Detection of Panton-Valentine leukocidin gene in *Staphylococcus aureus* by LightCycler PCR: clinical and epidemiological aspects. *Clin Microbiol Infect* **10**(10): 884-889.

Juhala RJ, Ford ME, Duda RL, Youlton A, Hatful GF & Hendrix RW (2000). Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. *J Mol Biol* **299**(1): 27-51.

Katayama Y, Ito T & Hiramatsu K (2000). A new class of genetic element, staphylococcus cassette chromosome *mec*, encodes methicillin resistance in *Staphylococcus aureus*. *Antimicrob Agents Chemother* **44**(6): 1549-1555.

Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, Santiago ML, Bibollet-Ruche F, Chen Y, Wain LV, Liegeois F, Loul S, Ngole EM, Bienvenue Y, Delaporte E, Brookfield JFY, Sharp PM, Shaw GM, Peeters M & Hahn BH (2006). Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1. *Science* **313**(5786): 523-526.

Kloos WE & Lambe DWJ (1991). Staphylococcus *in* A Balows (Hg) *Manual of Clinical Microbiology*. ASM Press, Washington DC, USA.

Knoop V & Müller K (2006). Gene und Stammbäume: Ein Handbuch zur molekularen Phylogenetik. Elsevier GmbH, München, Deutschland.

Knox R (1961). „Celbenin - resistant Staphylococci. *Br Med J* **1**(5219): 126.

Krziwanek K, Luger C, Sammer B, Stumvoll S, Stammler M, Sagel U, Witte W & Mittermayer H (2008). MRSA in Austria–an overview. *Clin Microbiol Infect* **14**(3): 250-259.

Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki K, Nagai Y, Lian J, Ito T, Kanamori M, Matsumaru H, Maruyama A, Murakami H, Hosoyama A, Mizutani-Ui Y, Takahashi NK, Sawano T, Inoue R, Kaito C, Sekimizu K, Hirakawa H, Kuhara S, Goto S, Yabuzaki J, Kanehisa M, Yamashita A, Oshima K, Furuya K, Yoshino C, Shiba T, Hattori M, Ogasawara N, Hayashi H & Hiramatsu K (2001). Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. *Lancet* **357**(9264): 1225-1240.

Kwan T, Liu J, DuBow M, Gros P & Pelletier J (2005). The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. *Proc Natl Acad Sci USA* **102**(14): 5174-5179.

Leplae R, Hebrant A, Wodak SJ & Toussaint A (2009). ACLAME: a CLAssification of Mobile genetic Elements. *J Bacteriol* **191**(11): 3462-3468.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R & 1000 Genome Project Data Processing Subgroup (2009). The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li H, Ruan J & Durbin R (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**(11): 1851-1858.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Yang H & Wang J (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20**(2): 265-272.

Librado P & Rozas J (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**(11): 1451-1452.

Lina G, Durand G, Berchich C, Short B, Meugnier H, Vandenesch F, Etienne J & Enright MC (2006). Staphylococcal chromosome cassette evolution in *Staphylococcus aureus* inferred from *ccr* gene complex sequence typing analysis. *Clin Microbiol Infect* **12**(12): 1175-1184.

Lindsay JA (2010). Genomic variation and evolution of *Staphylococcus aureus*. *Int J Med Microbiol* **300**(2-3): 98-103.

Lindsay, J (2008). *S. aureus* Evolution: Lineages and Mobile Genetic Elements (MGEs) *in* J Lindsay (Hg) *Staphylococcus: Molecular Genetics*. Caister Academic Press, Norfolk, UK.

Lindsay JA & Holden MT (2006). Understanding the rise of the superbug: investigation of the evolution and genomic variation of *Staphylococcus aureus*. *Funct Integr Genomics* **6**(3): 186-201.

Lindsay JA & Holden MT (2004). *Staphylococcus aureus*: superbug, super genome? *Trends Microbiol* **12**(8): 378-385.

Lowder BV, Guinane CM, Ben Zakour NL, Weinert LA, Conway-Morris A, Cartwright RA, Simpson AJ, Rambaut A, Nübel U & Fitzgerald JR (2009). Recent human-to-poultry host jump, adaptation, and pandemic spread of *Staphylococcus aureus*. *Proc Natl Acad Sci USA* **106**(46): 19545–19550.

Lowy FD (1998). *Staphylococcus aureus* infections. *N Engl J Med* **339**(8): 520-532.

Mantel N (1967). The detection of disease clustering and a generalized regression approach. *Cancer Res* **27**(2): 209-220.

Marchais A, Naville M, Bohn C, Bouloc P & Gautheret D (2009). Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. *Genome Res* **19**(6): 1084-1092.

Markoishvili K, Tsitlanadze G, Katsarava R, Glenn J & Sulakvelidze A (2002). A novel sustained-release matrix based on biodegradable poly(ester amide)s and impregnated with bacteriophages and an antibiotic shows promise in management of infected venous stasis ulcers and other poorly healing wounds. *Int J Dermatol* **41**(7): 453-458.

Marri PR, Hao W & Golding GB (2006). Gene gain and gene loss in *Streptococcus*: is it driven by habitat? *Mol Biol Evol* **23**(12): 2379-2391.

Mick V, Domínguez MA, Tubau F, Liñares J, Pujol M & Martín R (2010). Molecular characterization of resistance to Rifampicin in an emerging hospital-associated Methicillin-resistant *Staphylococcus aureus* clone ST228, Spain. *BMC Microbiol* **10**(68).

Monecke S, Coombs G, Shore AC, Coleman DC, Akpaka P, Borg M, Chow H, Ip M, Jatzwauk L, Jonas D, Kadlec K, Kearns A, Laurent F, O'Brien FG, Pearson J, Ruppelt A, Schwarz S, Scicluna E, Slickers P, Tan HL, Weber S & Ehricht R (2011). A field guide to pandemic, epidemic and sporadic clones of methicillin-resistant *Staphylococcus aureus*. *PLoS ONE* **6**(4): e17936.

Murtagh F (1984). Complexities of Hierarchic Clustering Algorithms: the state of the art. *Computation Stat Quarterly* **1**: 101-113.

Mwangi MM, Wu SW, Zhou Y, Sieradzki K, de Lencastre H, Richardson P, Bruce D, Rubin E, Myers E, Siggia ED & Tomasz A (2007). Tracking the in vivo evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. *Proc Natl Acad Sci USA* **104**(22): 9451-9456.

Nei M & Gojobori T (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**(5): 418-426.

Ning Z, Cox AJ & Mullikin JC (2001). SSAHA: a fast search method for large DNA databases. *Genome Res* **11**(10): 1725-1729.

Nübel U, Dordel J, Kurt K, Strommenger B, Westh H, Shukla SK, Žemličková H, Leblois R, Wirth T, Jombart T, Balloux F & Witte W (2010). A timescale for evolution, population expansion, and spatial spread of an emerging clone of methicillin-resistant *Staphylococcus aureus*. *PLoS Pathog* **6**(4): e1000855.

Nübel U, Roumagnac P, Feldkamp M, Song JH, Ko KS, Huang YC, Coombs G, Ip M, Westh H, Skov R, Struelens MJ, Goering RV, Strommenger B, Weller A, Witte W & Achtman M (2008). Frequent emergence and limited geographic dispersal of methicillin-resistant *Staphylococcus aureus*. *Proc Natl Acad Sci USA* **105**(37): 14130-14135.

Ou C-Y, Ciesielski CA, Myers G, Bandea CI, Luo C-C, Korber BTM, Mullins JI, Schochetman G, Berkelman RL, Economou AN, Witte JJ, Furman LJ, Satten GA, Maclnnes KA, Curran JW, Jaffe HW, Laboratory Investigation Group & Epidemiologic Investigation Group (1992). *Molecular epidemiology of HIV transmission in a dental practice. Science* **256**(5060): 1165-1171.

Pantucek R, Doskar J, Ruzickova V, Kasparek P, Oracova E, Kvardova V & Rosypal S (2004). Identification of bacteriophage types and their carriage in *Staphylococcus aureus*. *Arch Virol* **149**(9): 1689-1703.

Parkhill J & Wren BW (2011). Bacterial epidemiology and biology - lessons from genome sequencing. *Genome Biol* **12**(10): 230.

Pearson T, Busch JD, Ravel J, Read TD, Rhoton SD, U'Ren JM, Simonson TS, Kachur SM, Leadem RR, Cardon ML, Van Ert MN, Huynh LY, Fraser CM & Keim P (2004). Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing. *Proc Natl Acad Sci USA* **101**(37): 13536-13541.

Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG, Jacobs WRJ, Hendrix RW & Hatfull GF (2003). Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**(2): 171-182.

Pichon C & Felden B (2005). Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains. *Proc Natl Acad Sci USA* **102**(40): 14249-14254.

Pinho MG, de Lencastre H & Tomasz A (2001). An acquired and a native penicillin-binding protein cooperate in building the cell wall of drug-resistant staphylococci. *Proc Natl Acad Sci USA* **98**(19): 10886-10891.

Posada D & Crandall KA (1998). ModelTest: testing the model of DNA substitution. *Bioinformatics* **14**(9): 817-818.

Rambaut A & Drummond AJ (2007). „Tracer v1.5." *erhältlich unter* `http://tree.bio.ed.ac.uk/software/tracer/` [Stand: 03.08.2011].

Rambaut R (2011). „Path-O-Gen v1.3."*erhältlich unter* `http://tree.bio.ed.ac.uk/software/pathogen/` [Stand: 03.08.2011].

Robinson DA & Enright MC (2003). Evolutionary models of the emergence of methicillin-resistant *Staphylococcus aureus*. *Antimicrob Agents Chemother* **47**(12): 3926-3934.

Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH & Feil EJ (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J Theor Biol* **239**(2): 226-235.

Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J & Koonin EV (2002). Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res* **30**(19): 4264-4271.

Rohwer F & Edwards R (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184**(16): 4529-4535.

Ronaghi M, Uhlén M & Nyrén P (1998). A sequencing method based on real-time pyrophosphate. *Science* **281**(53750): 363-365.

Ronquist F & Huelsenbeck JP (2003). MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**(12): 1572-1574.

Roumagnac P, Weill FX, Dolecek C, Baker S, Brisse S, Chinh NT, Le TA, Acosta CJ, Farrar J, Dougan G & Achtman M (2004). Evolutionary history of *Salmonella* Typhi. *Science* **314**(5803): 1301-1304.

Rozen S & Skaletsky HJ (2000). Bioinformatics Methods and Protocols *in* S Misener & SA Krawetz (Hg) *Methods in Molecular Biology* **132**. Humana Press, Totowa, NJ, USA.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA & Barrell B (2000). Artemis: sequence visualization and annotation. *Bioinformatics* **16**(10): 944-945.

Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Ehrlich HA & Arnheim N (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**(4732): 1350-1354.

Sambrook J, Fritsch EF & Maniatis T (1989). Molecular Cloning: A Labratory Manual. Cold Spring Harbor Laboratory Press, New York, USA.

Sanger F, Nicklen S & Coulson AR (1977). DNA sequencing with chain-termination inhibitors. *Proc Natl Acad Sci USA* **75**(5463-5467.

Scaduto DI, Brown JM, Haaland WC, Zwickl DJ, Hillis DM & Metzker ML (2010). Source identification in two criminal cases using phylogenetic analysis of HIV-1 DNA sequences. *Proc Natl Acad Sci USA* **107**(50): 21242-21247.

Schmitt S (2006). Maximum Likelihood: Phylogenetische Anwendung. Online Publikation `http://www.informatik.uni-mainz.de/lehre/BioS/ml_biologie.pdf`. [Stand 23. September 2007]

Schuster SC (2008). Next-generation sequencing transforms today's biology. *Nat Methods* **5**(1): 16-18.

Shattock RJ & Moore JP (2003). Inhibiting sexual transmission of HIV-1 infection. *Nat Rev Microbiol* **1**(1): 25-34.

Shendure J & Ji H (2008). Next-generation DNA sequencing. *Nat Biotechnol* **26**(10): 1135-1145.

Short SM & Suttle CA (1999). Use of the polymerase chain reaction and denaturing gradient gel electrophoresis to study diversity in natural virus communities. *Hydrobiologia* **401**(0): 19-32.

Simon MN, Davis RW & Davidson N (1971). Heteroduplexes of DNA molecules of lambdoid phages: Physical mapping of their base sequence relationships by electron microscopy *in* AD Hershey (Hg) *The Bacteriophage*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

Smith JM, Smith NH, O'Rourke M & Spratt BG (1993). How clonal are bacteria? *Proc Natl Acad Sci USA* **90**(10): 4384-4388.

Sokal R & Michener C (1958). A statistical method for evaluating systematic relationships *in* University of Kansas Science Bulletin **38**: 1409-1438.

Spratt BG, Hanage WP & Feil EJ (2001). The relative contributions of recombination and point mutation to the diversification of bacterial clones. *Curr Opin Microbiol* **4**(5): 602-606.

Stangier K, Bauser C & Regenbogen J (2007). Praxisbericht: Next Generation DNA-Sequenzierung. Laborwelt **8**(3): 14-17.

Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W, Shi L (2011). Next-generation sequencing and its applications in molecular diagnostics. *Expert Rev Mol Diagn* **11**(3): 333-343.

Sulakvelidze A, Alavidze Z & Morris JG Jr. (2001). Bacteriophage therapy. *Antimicrob Agents Chemother* **45**(3): 649-659.

Tamura K, Dudley J, Nei M & Kumar S (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**(8): 1596-1599.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M & Kumar S (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol* **28**(10): 2731-2739.

Tenover FC & Goering RV (2009). Methicillin-resistant *Staphylococcus aureus* strain USA300: origin and epidemiology. *J Antimicrob Chemother* **64**(3): 441-446.

Thallinger G (2011). Vorlesung Bioinformatik - Next Generation Sequencing: Technologien und Anwendungen. Online Publikation `http://www.genome.tugraz.at/Bioinformatics/LectureNotes.pdf`. [Stand: 21. November 2011]

Theocharidis A, van Dongen S, Enright AJ & Freeman TC (2009). Network visualization and analysis of gene expression data using BioLayout Express^{3D}. *Nat Protoc* **4**(10): 1535-1550.

Valle J, Piriz S, de la Fuente R & Vadillo S (1991). Staphylococci isolated from healthy goats. *J Vet Med* Series B **38**(2): 81-89.

van Gils EJ, Hak E, Veenhoven RH, Rodenburg GD, Bogaert D, Bruin JP, van Alphen L & Sanders EA (2011). Effect of seven-valent pneumococcal conjugate vaccine on *Staphylococcus aureus* colonisation in a randomised controlled trial. *PLoS One* **6**(6): e20229.

Villaruz AE, Bubeck Wardenburg J, Khan BA, Whitney AR, Sturdevant DE, Gardner DJ, DeLeo FR & Otto M (2009). A point mutation in the agr locus rather than expression of the Panton-Valentine leukocidin caused previously reported phenotypes in *Staphylococcus aureus* pneumonia and gene regulation. *J Infect Dis* **200**(5): 724-734.

von Eiff C, Becker K, Machka K, Stammer H & Peters G (2001). Nasal carriage as a source of *Staphylococcus aureus* bacteremia. Study Group. *N Engl J Med* **344**(1): 11-16.

Wagner PL & Waldor MK (2002). Bacteriophage control of bacterial virulence. *Infect Immun* **70**(8): 3985-3993.

Westmoreland BC, Szybalski W & Ris H (1969). Mapping of deletions and substitutions in heteroduplex DNA molecules of bacteriophage lambda by electron microscopy. *Science* **163**(3873): 1343 - 1348.

Wilgenbusch JC & Swofford D (2003). Inferring evolutionary trees with PAUP*. *Curr Protoc Bioinformatics* **Chapter 6:** Unit 6.4.

Willems RJ, Hanage WP, Bessen DE & Feil EJ (2011). Population biology of Gram-positive pathogens: high-risk clones for dissemination of antibiotic resistance. *FEMS Microbiol Rev* **35**(5): 872-900.

Witte W, Cuny C, Braulke C & Heuk D (1994). Clonal dissemination of two MRSA strains in Germany. *Epidemiol Infect* **113**(1): 67-73.

Witte W, Cuny C, Klare I, Nübel U, Strommenger B & Werner G (2008). Emergence and spread of antibiotic resistant Gram positive bacterial pathogens. *Int J Med Microbiol* **298**(5-6): 365-377.

Woese CR, Kandler O & Wheelis ML (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* **87**(12): 4576-4579.

Wommack KE & Colwell RR (2000). Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* **64**(1): 69-114.

Wright A, Hawkins CH, Änggård EE & Harper DR (2009). A controlled clinical trial of a therapeutic bacteriophage preparation in chronic otitis due to antibiotic-resistant *Pseudomonas aeruginosa*; a preliminary report of efficacy. *Clin Otolaryngol* **34**(4): 349-357.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'Haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng JF, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D, Rubin EM, Kyrpides NC, Klenk HP & Eisen JA (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**(7276): 1056-1060.

Yang L, Jelsbak L, Marvig RLL, Damkiær S, Workman CT, Rau MHH, Hansen SKK, Folkesson A, Johansen HKK., Ciofu O, Høiby N, Sommer MO & Molin S (2011). Evolutionary dynamics of bacteria in a human host environment. *Proc Natl Acad Sci USA* **108**(18): 7481-7486.

Anhang

Tabelle A.1: Primer für die Verifizierung von homoplastischen SNPs in CC5.

Name	Sequenz (5' → 3')	Richtung	Produktgröße (bp)
1.14603_f	tag ggt agg ggg aga gga tg	vorwärts	230
1.14603_r	tta gac acc agt ttg tct gag gt	rückwärts	
3.229185_f	ctc gcg cag cat cta aga a	vorwärts	294
3.229185_r	cct ttg att cct gtc caa gc	rückwärts	
4.885644_f	tcg cgg tat tgt tca ttt tg	vorwärts	538
4.885644_r	tgt cgc ttt atc cac cat ca	rückwärts	
5.2107968_f	tga acc ttc agg tgc aag tag	vorwärts	470
5.2107968_r	caa att ttg taa tat cgt ctt gtg g	rückwärts	
6.2240290_f	tgg gct agt ttc ctc ttg ga	vorwärts	540
6.2240290_r	aaa ggg ctc acg cta ttt tat g	rückwärts	
7.2308472_f	tca tcg aaa gtc cac ctc ct	vorwärts	391
7.2308472_r	ttt tcg tgg aga agt cta tca ct	rückwärts	

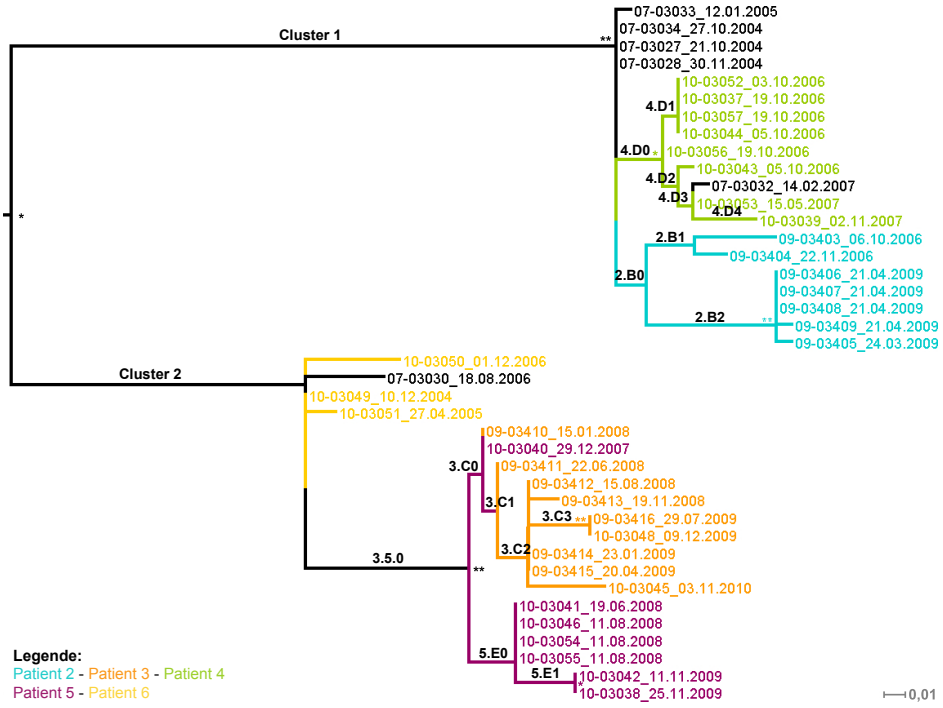

Abbildung A.1: „*Maximum Likelihood*"-Baum der ST225 Patienten-Isolate. Die Markierungen entsprechen der Benennung der Linien in Tabelle 6.2.

Tabelle A.2: Linienspezifische SNPs in den Patienten.

Patient	Linie	SNP Bezeichnung	Position in 04-02981	Ref	SNP	AA	Typ	Locus Tag in 04-02981	Gen	Produkt
2	2.B0	2.B0_1	491274	T	C	G→G	s	SA2981_0447	gltB	Glutamate synthase [NADPH] large chain
2	2.B0	2.B0_2	1022862	C	T	-	i	-	-	-
2	2.B0	2.B0_3	2715372	T	C	N→S	ns	SA2981_2559	*bsaA*	Glutathione peroxidase family protein
2	2.B1	2.B1_1	215382	A	G	S→S	s	SA2981_0190	*glcA*	PTS system, glucose-specific IIA-IIC component
2	2.B1	2.B1_2	746339	A	G	E→E	s	SA2981_0677	*fruA*	PTS system, fructose-specific IIA-IIC component
2	2.B1	2.B1_3	1432402	T	C	S→S	s	SA2981_1346	-	ABC transporter ATP-binding protein
2	09-03403	spezifisch	560517	G	A	G→D	ns	SA2981_0500	*clpC*	ATP-dependent Clp protease, ATP-binding subunit ClpC / Negative regulator of genetic competence clcC/mecB
2	09-03403	spezifisch	745706	G	A	M→I	ns	SA2981_0677	*fruA*	PTS system, fructose-specific IIA-IIC component
2	09-03403	spezifisch	1162591	A	G	-	i	-	-	-
2	09-03403	spezifisch	1202981	G	A	G→R	ns	SA2981_1146	-	protein of unknown function
2	09-03403	spezifisch	2695384	G	T	E→*	ns	SA2981_2544	*fda*	Fructose-bisphosphate aldolase class I
2	09-03404	spezifisch	271190	T	C	D→G	ns	SA2981_0233	-	Glutaryl-CoA dehydrogenase
2	09-03404	spezifisch	1672871	C	T	M→I	ns	SA2981_1564	*accC*	putative biotin carboxylase subunit of acetyl-CoA carboxylase
2	09-03404	spezifisch	1929510	A	G	I→T	ns	SA2981_1799	-	DNA double-strand break repair rad50 ATPase
2	2.B2	2.B2_1	256344	A	G	V→A	ns	SA2981_0224	-	Two-component sensor histidine kinase
2	2.B2	2.B2_2	1013365	T	C	N→N	s	SA2981_0960	-	GTP pyrophosphokinase
2	2.B2	2.B2_3	1532404	T	C	D→G	ns	SA2981_1413	-	tRNA nucleotidyltransferase
2	2.B2	2.B2_4	2138186	A	G	L→*	ns	SA2981_2005	*sigB*	RNA polymerase sigma factor SigB
2	2.B2	2.B2_5	2275547	G	A	P→S	ns	SA2981_2126	-	cell surface hydrolase (putative)
2	2.B2	2.B2_6	2397090	A	G	L→L	s	SA2981_2268	-	N-acetyl-L,L-diaminopimelate deacetylase
2	2.B2	2.B2_7	2398536	T	C	A→A	s	SA2981_2269	*hutI*	Imidazolonepropionase
2	2.B2	2.B2_8	2692552	G	A	A→V	ns	SA2981_2542	-	Aminotransferase
2	09-03405	spezifisch	727406	C	T	M→I	ns	SA2981_0658	-	hypothetical protein
2	09-03409	spezifisch	1425252	C	T	D→N	ns	SA2981_1339	*phoU*	Phosphate transport system regulatory protein PhoU
2	09-03409	spezifisch	2101163	A	C	D→E	ns	SA2981_1982	*scrB*	Sucrose-6-phosphate hydrolase
3	3.C0	3.C0	972013	T	C	R→R	s	SA2981_0922	-	ATP-dependent nuclease, subunit A
3	3.C1	3.C1_1	1344275	G	T	-	i	-	-	-
3	3.C1	3.C1_2	2429189	T	C	H→R	ns	SA2981_2299	*hrtB*	Heme efflux system permease HrtB
3	3.C2	3.C2_1	821597	A	G	-	i	-	-	-
3	3.C2	3.C2_2	950452	C	T	V→I	ns	SA2981_0907	*mnhA*	Na(+) H(+) antiporter subunit A
3	3.C2	3.C2_3	959705	G	T	R→S	ns	SA2981_0914	*glpQ*	putative glycerophosphoryl diester phosphodiesterase
3	10-03045	spezifisch	9120	G	A	V→I	ns	SA2981_0006	*gyrA*	DNA gyrase subunit A
3	10-03045	spezifisch	121473	C	T	-	i	-	-	-
3	10-03045	spezifisch	170509	C	T	FR→F*	s/ns	SA2981_0153/SA2981_0154	*cap8D/cap8E*	Capsular polysaccharide synthesis enzyme Cap8D/Cap8E
3	10-03045	spezifisch	497218	A	G	K→E	ns	SA2981_0450	*treC*	Trehalose-6-phosphate hydrolase
3	10-03045	spezifisch	2198631	G	A	R→C	ns	SA2981_2067	*ctrA*	CTP synthase
3	10-03045	spezifisch	2733631	T	G	-	i	-	-	-

Abkürzungen: ns: nicht-synonym, s: synonym i: intergenisch

Tabelle A.2: Linienspezifische SNPs in den Patienten.

Patient	Linie	SNP Bezeichnung	Position in 04-02981	Ref	SNP	AA	Typ	Locus Tag in 04-02981	Gen	Produkt
3	3.C3	3.C3_1	257514	C	T	-	i	-	-	-
3	3.C3	3.C3_2	514791	G	C	A→G	ns	SA2981_0462	-	Predicted O-methyltransferase
3	3.C3	3.C3_3	857398	A	G	-	i	-	-	-
3	3.C3	3.C3_4	1152480	G	A	G→D	ns	SA2981_1095	*pheT*	Phenylalanyl-tRNA synthetase beta chain
3	3.C3	3.C3_5	1942282	G	A	A→V	ns	SA2981_1811	-	tRNA (cytosine34-2-O-)-methyltransferase
3	3.C3	3.C3_6	2035991	G	A	-	i	-	-	-
3	3.C3	3.C3_7	2682010	T	G	W→G	ns	SA2981_2532	-	Predicted acyl esterase/dipeptidyl-peptidase
3	09-03413	spezifisch	1266674	T	C	I→T	ns	SA2981_1203	*sucC*	Succinyl-CoA ligase [ADP-forming] beta chain
3	09-03413	spezifisch	2802798	C	T	P→L	ns	SA2981_2633	-	hypothetical protein
4	4.D0	4.D0_1	286885	C	T	A→V	ns	SA2981_0244	-	Predicted galactitol operon regulator (Transcriptional antiterminator), BglG family / PTS system, mannitol/fructose-specific IIA component
4	4.D0	4.D0_2	1378142	G	C	L→L	s	SA2981_1301	sbcC	Exonuclease SbcC
4	4.D0	4.D0_3	1831533	T	G	E→D	ns	SA2981_1705	-	metallo-beta-lactamase family protein
4	4.D1	4.D1_1	331868	G	A	Q→Q	s	SA2981_0287	-	FtsK/SpoIIIE family protein, putative secretion system component EssC/YukA
4	4.D1	4.D1_2	2154145	C	T	-	i	-	-	-
4	4.D2	4.D2_1	2609976	G	A	Q→*	ns	SA2981_2459	-	Oxygen-insensitive NAD(P)H nitroreductase / Dihydropteridine reductase
4	10-03043	spezifisch	2430929	G	A	A→T	ns	SA2981_2301	-	Sensor histidine kinase colocalized with HrtAB transporter
4	4.D3	4.D3_1	1745586	C	T	R→H	ns	SA2981_1638	*rplT*	50S ribosomal protein L20
4	4.D4	4.D4_1	578083	A	C	N→H	ns	SA2981_0519	*rpoB*	DNA-directed RNA polymerase beta subunit
4	4.D4	4.D4_2	734817	C	A	R→S	ns	SA2981_0665	-	hypothetical protein
4	4.D4	4.D4_3	844699	G	T	D→*	ns	SA2981_0764	-	hypothetical protein
4	4.D4	4.D4_4	2140283	G	A	A→V	ns	SA2981_2008	*rsbU*	Serine phosphatase RsbU, regulator of sigma subunit
4	4.D4	4.D4_5	2592923	A	G	C→R	ns	SA2981_2444	-	Transcriptional regulator, MerR family
5	5.E0	5.E0_1	101221	G	A	K→K	s	SA2981_0094		hypothetical protein
5	5.E0	5.E0_2	809969	T	G	D→E	ns	SA2981_0734	*yfbR*	Nucleotidase YfbR, HD superfamily
5	5.E0	5.E0_3	383020	A	G	-	i	-	-	-
5	5.E0	5.E0_4	1269498	G	A	R→Q	ns	SA2981_1206	*fmhC*	FmhC protein of FemAB family
5	5.E0	5.E0_5	1608140	G	A	N→N	s	SA2981_1493	-	Glycine dehydrogenase [decarboxylating] (glycine cleavage system P2 protein)
5	5.E0	5.E0_6	2186192	G	A	N→N	s	SA2981_2055	-	Low molecular weight protein tyrosine phosphatase
5	5.E1	5.E1_1	479421	G	A	-	i	-	-	-
5	5.E1	5.E1_2	1239421	C	T	T→I	ns	SA2981_1179	*engC*	Ribosome small subunit-stimulated GTPase EngC
5	5.E1	5.E1_3	1785094	G	A	L→F	ns	SA2981_1667	*ald*	Alanine dehydrogenase
5	5.E1	5.E1_4	2263385	G	A	A→V	ns	SA2981_2116	-	Hypothetical protein
5	10-03038	spezifisch	2602215	C	T	-	i	-	-	-
6	10-03050	spezifisch	263771	G	T	-	i	-	-	-
6	10-03050	spezifisch	677408	C	T	-	i	-	-	-
6	10-03050	spezifisch	1077348	T	C	Y→H	ns	SA2981_1021	*purK*	Phosphoribosylaminoimidazole carboxylase ATPase subunit
6	10-03050	spezifisch	1195193	G	A	A→T	ns	SA2981_1140	*mraY*	Phospho-N-acetylmuramoyl-pentapeptide-transferase
6	10-03050	spezifisch	1272084	G	A	E→E	s	SA2981_1208	*topA*	DNA topoisomerase I

Abkürzungen: ns: nicht-synonym, s: synonym, i: intergenisch

Tabelle A.2: Linienspezifische SNPs in den Patienten.

Patient	Linie	SNP Bezeichnung	Position in 04-02981	Ref	SNP	AA	Typ	Locus Tag in 04-02981	Gen	Produkt
6	10-03050	spezifisch	1845601	A	G	I→T	ns	SA2981_1714	*sasC*	Predicted cell-wall-anchored protein SasC (LPXTG motif)
6	10-03050	spezifisch	2139861	A	T	L→I	ns	SA2981_2008	*rsbU*	Serine phosphatase RsbU, regulator of sigma subunit
6	10-03050	spezifisch	2449095	G	C	-	i	-	-	-
6	10-03051	spezifisch	1949609	T	A	-	i	-	-	-
6	10-03051	spezifisch	2334588	C	T	V→V	s	SA2981_2200	*fmhB*	FmhB protein of FemAB family involved in peptidoglycan interpeptide biosynthesis
6	10-03051	spezifisch	2590067	C	A	A→S	ns	SA2981_2441	*gntP*	Gluconate permease
6	10-03050	spezifisch	1272084	G	A	E→E	s	SA2981_1208	*topA*	DNA topoisomerase I
6	10-03050	spezifisch	1845601	A	G	I→T	ns	SA2981_1714	*sasC*	Predicted cell-wall-anchored protein SasC (LPXTG motif)
6	10-03050	spezifisch	2139861	A	T	L→I	ns	SA2981_2008	*rsbU*	Serine phosphatase RsbU, regulator of sigma subunit
6	10-03050	spezifisch	2449095	G	C	-	i	-	-	-
6	10-03051	spezifisch	1949609	T	A	-	i	-	-	-
6	10-03051	spezifisch	2334588	C	T	V→V	s	SA2981_2200	*fmhB*	FmhB protein of FemAB family involved in peptidoglycan interpeptide biosynthesis
6	10-03051	spezifisch	2590067	C	A	A→S	ns	SA2981_2441	*gntP*	Gluconate permease

Abkürzungen: ns: nicht-synonym, s: synonym, i: intergenisch

Tabelle A.3: Primer für die Verifizierung von SNPs in den ST225 Patientenisolaten. Die Linienbezeichnung entspricht der Markierung in Abbildung A.1.

Patient	Linie	SNP Bezeichnung	Primer Name	Sequenz (5'→3')	Richtung	Produktgröße
2	2.B	2.B1	2.B1_f	cgt ttt gta gcg gta aca tca	vorwärts	274
2	2.B	2.B1	2.B1_r	caa tga cgg cat aca aac ctt	rückwärts	
2	2.B	2.B2	2.B2_f	tcg ctg gtt taa gaa tga tgt g	vorwärts	483
2	2.B	2.B2	2.B2_r	ggc att aga ctt gga gtc acc	rückwärts	
2	2.B	2.B3	2.B3_f	ttc tat ttc atc cgg cgt tg	vorwärts	355
2	2.B	2.B3	2.B3_r	gtg ggt gat tgt gat tca gc	rückwärts	
2	2.B	2.B4	2.B4_f	aaa aat acc cct cga ttt caa	vorwärts	490
2	2.B	2.B4	2.B4_r	agc gat gaa cta acc gct ga	rückwärts	
2	2.B	2.B5	2.B5_f	tca ctg cac cat cct ttg aa	vorwärts	185
2	2.B	2.B5	2.B5_r	tca acg aca ggc aca aac at	rückwärts	
2	2.B	2.B6	2.B6_f	gtgctggcgtcatttcagat	vorwärts	597
2	2.B	2.B6	2.B6_r	ggcacagtgcgtagttttga	rückwärts	
2	2.B	2.B7	2.B7_f	ttttacaaaccacaaaagctctaaa	vorwärts	400
2	2.B	2.B7	2.B7_r	tgcgagaggtatgcttgataa	rückwärts	
2	2.B	2.B8	2.B8_f	cgctattaccaaccttgtcca	vorwärts	274
2	2.B	2.B8	2.B8_r	gaggtaaaggtctgagcattgg	rückwärts	
3	3.C1	3.C1_0	3.C1_0_f	tggcgccacgagtaaagta	vorwärts	696
3	3.C1	3.C1_0	3.C1_0_r	tcgcacttctatagcttcctgtt	rückwärts	
3	3.C1	3.C1_1	3.C1_1_f	cattacataaaaaccttaggtgctc	vorwärts	631
3	3.C1	3.C1_1	3.C1_1_r	gagtttttccgctgtttctaaag	rückwärts	
3	3.C1	3.C1_2	3.C1_2_f	tgaggcacctaaaatcgctac	vorwärts	272
3	3.C1	3.C1_2	3.C1_2_r	caatcgttctaagtgcatttttc	rückwärts	
3	3.C2	3.C2_1	3.C2_1_f	tcatttgtgcgtttcaatgag	vorwärts	933
3	3.C2	3.C2_1	3.C2_1_r	ggatggatttccttgttcca	rückwärts	
3	3.C2	3.C2_2	3.C2_2_f	tttttcacctcgttaccttgc	vorwärts	397
3	3.C2	3.C2_2	3.C2_2_r	aagcacccgacttagcattg	rückwärts	
3	3.C2	3.C2_3	3.C2_3_f	gcgaaatttgtaaagacaccatc	vorwärts	289
3	3.C2	3.C2_3	3.C2_3_r	ttcaatcattttctgacgaaagtt	rückwärts	
4	4.D0	4.D0_1	4.D0_1_f	catgagatgcaacaggtatttga	vorwärts	476
4	4.D0	4.D0_1	4.D0_1_r	ttctctgtcgatgactgcatct	rückwärts	
4	4.D0	4.D0_2	4.D0_2_f	tggtaaacagccgatgtcag	vorwärts	686
4	4.D0	4.D0_2	4.D0_2_r	tttcaggttgttgctctttcaa	rückwärts	
4	4.D0	4.D0_3	4.D0_3_f	agtaagattggatgtggctca	vorwärts	495
4	4.D0	4.D0_3	4.D0_3_r	tggcatgagtttattgcacct	rückwärts	
4	4.D1	4.D1_1	4.D1_1_f	ggacatcttgatgaagcgatta	vorwärts	998
4	4.D1	4.D1_1	4.D1_1_r	atatcatcggtcggttcacg	rückwärts	
4	4.D1	4.D1_2	4.D1_2_f	gcactagcgtcagcaaaaag	vorwärts	1995
4	4.D1	4.D1_2	4.D1_2_r	gcaaaggccgtaatggtaaa	rückwärts	
4	4.D2	4.D2_1	4.D2_1_f	ccagactaaaaccttccatcg	vorwärts	599
4	4.D2	4.D2_1	4.D2_1_r	gcgaattgcaaacggattac	rückwärts	
4	4.D3	4.D3_1	4.D3_1_f	gacatcgtataaaactcctcacttt	vorwärts	768
4	4.D3	4.D3_1	4.D3_1_r	ctaagttgaaaaacggacgtaaa	rückwärts	
4	4.D3	4.D3_3	4.D3_3_f	tggattccaagatggttgct	vorwärts	816
4	4.D3	4.D3_3	4.D3_3_r	ggtgcaccgcaattagattt	rückwärts	
4	4.D4	4.D4_1	4.D4_1_f	agttcgttcaccatccgttt	vorwärts	665
4	4.D4	4.D4_1	4.D4_1_r	tcgtacgaccttcatcatcg	rückwärts	
4	4.D4	4.D4_2	4.D4_2_f	ccggaatctgttctacgtttg	vorwärts	696
4	4.D4	4.D4_2	4.D4_2_r	tggtgaaggacgatgaatga	rückwärts	
4	4.D4	4.D4_3	4.D4_3_f	gcaaaacttttatatgagcagtcg	vorwärts	690
4	4.D4	4.D4_3	4.D4_3_r	tttttcgcaataaacaactaccc	rückwärts	
4	4.D4	4.D4_5	4.D4_5_f	aaacattgtgacgaacatattttga	vorwärts	399
4	4.D4	4.D4_5	4.D4_5_r	gaaatcgttaaaggctttggtt	rückwärts	
4	4.D4	4.D4_6	4.D4_6_f	gaccacccctcttcgatgta	vorwärts	242
4	4.D4	4.D4_6	4.D4_6_r	tttatgtcctaactgccagcat	rückwärts	
5	5.E1	5.E1_1	5.E_1_f	gcaacatcgtattacacattagctg	vorwärts	590
5	5.E1	5.E1_1	5.E_1_r	ttgcgattgaccgtaaatga	rückwärts	
5	5.E1	5.E1_2	5.E_2_f	tgttgagattgcggtagtttt	vorwärts	827
5	5.E1	5.E1_2	5.E_2_r	gaatccaggtgtgtctgcaa	rückwärts	
5	5.E1	5.E1_3	5.E_3_f	tgttacaagtcttggcgcttt	vorwärts	700
5	5.E1	5.E1_3	5.E_3_r	tttatcaccgagtggtgtgc	rückwärts	
5	5.E1	5.E1_4	5.E_4_f	cgtttcacctattcgcaattc	vorwärts	655
5	5.E1	5.E1_4	5.E_4_r	catggcacacgttaacataaa	rückwärts	

Abkürzungsverzeichnis

A-E

AA	Aminosäure
AIC	Akaike Informationskriterium
AICc	korrigiertes Akaike Informationskriterium
APS	Adenosin-5'-phosphosulfat
ATP	Adenosintriphosphat
bp	Basenpaare
BS	Bootstrap-Werte
CA-MRSA	engl. „*Community Acquired*“ MRSA
CC	klonaler Komplex; engl. *„Clonal Complex“*
CDS	kodierende Sequenz; engl. „*CoDing Sequence*“
dN	Verhältnis nicht-synonymer SNPs zu nicht-synonymen Seiten in einer kodierenden Region
dNTP	Desoxynukleotid
ddNTP	Didesoxynukleotid
dS	Verhältnis synonymer SNPs zu synonymen Seiten in einer kodierenden Region
ESS	engl. „*Effective Sample Size*“
et al.	*et alii* (und andere)
E-Wert	gibt die erwartete Anzahl der Hits an, deren Score mindestens so groß ist wie der beobachtete (je kleiner, desto besser)

F-J

HA-MRSA	Krankenhaus-assoziierte MRSA; engl. „*Hospital Acquired MRSA*“
HI	Homoplasie-Index
HIV, HI-Virus	Humanes Immundefizienz-Virus
IS	Insertionselement

K-O

kb	Kilobasenpaare (1 x 10^3)
LA-MRSA	engl. „*Livestock Associated*“ MRSA
LR-ELW	Statistische Unterstützung von Knoten in einem phylogenetischen Stammbaum (approximierte Bootstrap-Werte); engl. „*Local Rearrangements Expected-Likelihood Weights*“
Mb	Megabasenpaare (1 x 10^6)
MCMC	engl. „*Markov Chain Monte Carlo*“
MCMCMC	engl. „*Metropolis-coupled Markov Chain Monte Carlo*“

ML	Maximum Likelihood
MLST	Multi-Lokus-Sequenz-Typisierung
MRSA	Methicillin-Resistente *Staphylococcus aureus*
MSSA	Methicillin-Sensible *Staphylococcus aureus*
n.a.	Information nicht verfügbar; engl. „*not available*“
ncRNA	nicht-kodierende RNA; engl. „*Non Coding RNA*“
NGS	engl. „*Next-Generation Sequencing*“
ns	nicht-synonym
ORF	offener Leserahmen; engl. „*Open Reading Frame*“

P-T

P_e	Fehlerwahrscheinlichkeit einer identifizierten Base; engl. „*base calling error probability*“
PBP	Penicillin-Bindeprotein
PCR	Polymerase-Kettenreaktion; engl. „*Polymerase Chain Reaction*“
PP	engl. „*Posterior Probability*“
PP_i	anorganisches Pyrophosphat
PVL	Panton-Valentine Leukozidin
Q_{PHRED}	PHRED-Qualitätswert
r/m	relative Likelihood, dass ein Polymorphismus eher durch Rekombination als durch Mutation entstanden ist
s	synonym
S. aureus	*Staphylococcus aureus*
Sa*int*	Integrasegruppe der *S. aureus* Prophagen
SaPI	Pathogenitätsinsel
SCC*mec*	engl. „*Staphylococcal Chromosomal Cassette mec*“
SLV	Einzelnukleotidvarianten; engl. „*Single Locus Variants*“
SNP	Einzelnukleotidpolymorphismus; engl. *„Single Nucleotide Polymorphism“*
SSS	engl. „*Staphylococcal Scalded Skin Syndrom*“
ST	Sequenztyp; engl. „*Sequence Type*“
Tn	Transposon
TMRCA	engl. „*Time to Most Recent Common Ancestor*
TSS	Toxisches Schock Syndrom

V-Z

VISA	Vancomycin-Intermediär-Sensible *S. aureus*
VRSA	Vancomycin-Resistente *S. aureus*
VSSA	Vancomycin-sensitive *S. aureus*
UPGMA	engl. „*Unweighted Pair Group Method with Arithmetic mean*“
XML	engl. „*eXtensible Markup Language*“; Auszeichnungssprache zur Darstellung hierarchisch strukturierter Daten in Form von Textdaten

Gennamen

aur	Aureolysin
chp	Chemotaxis Inhibitor Protein
ear	putatives
	β–Lactamase Protein
egc	Enterotoxin Cluster
ermA	Erythromycin-Resistenz
groEL	60 kDa Chaperonin
guaA	GMP Synthase
lpl	Lipoprotein Gen Cluster
lukDE	Komponenten des Leukozidin DE Toxins
sak	Staphylokinase
scn	Staphylokokken Komplement Inhibitor
sea	Enterotoxin A
seb	Enterotoxin B
sec1-3	Enterotoxin C1-3
sed	Enterotoxin D
see	Enterotoxin E
seg	Enterotoxin G
seh	Enterotoxin H
sei	Extrazelluläres Enterotoxin Typ I
sak	Staphylokinase
sek	Enterotoxin K
sel	Enterotoxin L
sem	Enterotoxin M
seo	Enterotoxin O
sep	Enterotoxin P
set	Staphylokokken Exotoxin Gen Cluster
spc	Streptomycin 3'-adenylyltransferase
spl	Staphylokokken Serin Protease Cluster
sufB	FeS assembly Protein
tetM	Tetracycline Resistenz Protein
tst-1	„*toxic shock syndrome*"-Toxin-1
yent1	Enterotoxin Yent1
yent2	Enterotoxin Yent2

Abbildungsverzeichnis

Tabellenverzeichnis

Danksagung

An dieser Stelle möchte ich all denen danken, die Anteil an der Anfertigung dieser Dissertation hatten.
DANKE an:

Prof. Dr. Dieter Jahn für die Ermöglichung der Durchführung dieser Arbeit und die Begleitung des Promotionsverfahrens als Mentor.

PD Dr. Ulrich Nübel für die hervorragende Betreuung, Unterstützung und Diskussionsbereitschaft. Danke für die Chance, die ich bekommen habe!

Prof. Dr. Wolfgang Witte für die Förderung, Betreuung und das Interesse am Fortlauf der Arbeit.

PD Dr. Max Schobert für die Bereitschaft als Fachprüfer der Promotionskommission beizusitzen und den Vorsitz zu übernehmen.

Das gesamte RKI Wernigerode und besonders FG13 für die freundliche Aufnahme, die entgegengebrachte Hilfe und viele informative, produktive und lustige Tage!

Annette Weller, Heike Illiger und Mike Henkel für Unterstützung im Labor!

Meine Lieblingsinformatiker Wojtek Dabrowski vom RKI Berlin, Jochen Blom vom CeBiTec Bielefeld und Stefan Pitz für allerschnellste Hilfe bei Computer-, Linux-, Programmier- und Softwareproblemen! Code Monkey!

Angela Cullik, Anne Gutsche, Jenny Laverde Gomez und Henning Zaiß für ihre Freundschaft. Ich bin froh, dass ich Euch kennenlernen durfte! Danke für Gespräche und Diskussionen, Aufmunterungen, Ausflüge, Grill- und Tanzabende auf der Dachterasse und AFAFLEI!

Kevin Kurt und Christoph „Schäfchen“ Eller für das gemeinsame Durchstehen der letzten Monate, ausgiebige Büro- und Konzertbesuche und den Flusen in Euren Köpfen! Stay lucky!

Anne-Kathrin Exner, Simon Hoff, Andreas Josch, Martin „Pogo“ Stock und Monika Wirtz für so viele schöne gemeinsame Jahre! Danke, dass Ihr immer für mich da seid! Won’t forget these days!

Meine Eltern Michaela und Wolfgang Dordel und meine Schwester Carolin Lehmann. Danke für Eure immerwährende Unterstützung, ohne die ich nicht so weit gekommen wäre!